EXHIBIT A

FUNCTION

BestFit makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the *local homology* algorithm of Smith and Waterman.

DESCRIPTION

- BestFit inserts gaps to obtain the optimal alignment of the best region of similarity between two sequences, and then displays the alignment in a format similar to the output from Gap. The sequences can be of very different lengths and have only a small segment of similarity between them. You could take a short RNA sequence, for example, and run it against a whole mitochondrial genome.

SEARCHING FOR SIMILARITY

BestFit is the most powerful method in the Wisconsin Sequence Analysis Package™ for identifying the best region of similarity between two sequences whose relationship is unknown.

EXAMPLE

The sequence gamma.seq contains an Alu family sequence somewhere in the first 500 bases. alu.seq contains a generic human Alu family repeat. The two sequences are aligned and the best segment of similarity is found with BestFit.

```
% bestfit

BESTFIT of what sequence 1 ?  gamma.seq

                Begin (* 1 *) ?
              End (* 11375 *) ?  500
            Reverse (* No *) ?

to what sequence 2 (* gamma.seq *) ?  alu.seq

                Begin (* 1 *) ?
              End (*   207 *) ?
            Reverse (* No *) ?

What is the gap creation penalty (* 5.00 *) ?

What is the gap extension penalty (* 0.30 *) ?

What should I call the paired output display file (* gamma.pair *)

Aligning ..........-..

        Gaps:      3
     Quality: 129.3
Quality Ratio: 0.625
 % Similarity: 84.466
      Length:   209

%
```

OUTPUT

Here is the output file. Notice how BestFit finds and displays only the best segments of similarity:

```
 BESTFIT of: gamma.seq  check: 6474  from: 1  to: 500

Human fetal beta globins G and A gamma
from Shen, Slightom and Smithies,  Cell 26; 191-203.
Analyzed by Smithies et al. Cell 26; 345-353.

 to: alu.seq  check: 4238  from: 1  to: 207

HSREP2 from the EMBL data library
Human Alu repetitive sequence located near the insulin gene
Dhruva D.R., Shenk T., Subramanian K.N.; "Integration in vivo into
Simian virus 40 DNA of a sequence that resembles a certain family of
genomic interspersed repeated sequences"; Proc. Natl. Acad. Sci. USA
77:4514-4518(1980).  . . .

 Symbol comparison table: Gencoredisk:[Gcgcore.Data.Rundata]Swgapdna.Cmp
 CompCheck: 5234

        Gap Weight:  5.000      Average Match:  1.000
     Length Weight:  0.300   Average Mismatch: -0.900

           Quality:  129.3             Length:    209
             Ratio:  0.625               Gaps:      3
Percent Similarity:  84.466   Percent Identity: 84.466

gamma.seq x alu.seq       June 20, 1994 15:15

  137 AGACCAACCTGGCCAACATGGTGAAATCCCATCTCTAC.AAAAATACAAA 185
      |||||| ||:|||||||||||||||||| _|||||||||| _|||||||||
    1 AGACCAGCCTGGCCAACATGGTGAAACTCCATCTCTACTGAAAATACAAA 50

  186 AATTAGACAGGCATGATGGCAAGTGCCTGTAATCCCAGCTACTTGGGAGG 235
      |||||| |||||||| ||  ||||||| ||||||||||||||| |||||
   51 AATTAGCCAGGCATGGTGATGCGTGCCTGGAATCCCAGCTACTTAGGAGG 100

  236 CTGAGGAAGGAGAATTGCTTGAACCTGGAAGGCAGGAGTTGCAGTGAGCC 285
      ||||| || |||||_|||_|||| _| ||| _| |_ |||||||||||||
  101 CTGAGACAGAAGAATCCCTTAAACCAAG.AGGTGGAGGTTGCAGTGAGCC 149

  286 GAGATCATACCACTGCACTCCAGCCTGGGTGACAGAACAAGACTCTGTCT 335
      |||||| || |||||||||||||| |||||||| | |||||| |||
  150 GAGATCGCACGGCTGCACTCCAGCCT.GGTGACAGAGCGAGACTCCATCT 198

  336 CAAAAAAAA 344
      | | . .
  199 CAAAAAAAA 207
```

RELATED PROGRAMS

When you want an alignment that covers the whole length of both sequences, use Gap. When you are trying to find only the best segment of similarity between two sequences, use BestFit. PileUp creates a multiple sequence alignment of a group of related sequences, aligning the whole length of all sequences. DotPlot displays the entire surface of comparison for a comparison of two sequences. GapShow displays the pattern of differences between two aligned sequences. PlotSimilarity plots the average similarity of two or more aligned sequences at each position in the alignment. Pretty displays alignments of several sequences. LineUp is an editor for editing multiple sequence alignments. CompTable helps generate scoring matrices for peptide comparison.

ALGORITHM

BestFit uses the *local homology* algorithm of Smith and Waterman (Advances in Applied Mathematics 2; 482-489 (1981)) to find the best segment of similarity between two sequences. BestFit reads a scoring matrix that contains values for every possible GCG symbol match (see the LOCAL DATA FILES topic below). The program uses these values to construct a path matrix that represents the entire surface of comparison with a score at every position for the best possible alignment to that point. The *quality* score for the best alignment to any point is equal to the sum of the scoring matrix values of the matches in that alignment, less the gap creation penalty times the number of gaps in that alignment, less the gap extension penalty times the total length of all gaps in that alignment. The gap creation and gap extension penalties are set by you. If the best path to any point has a negative value, a zero is put in that position.

After the path matrix is complete, the highest value on the surface of comparison represents the end of the best region of similarity between the sequences. The best path from this highest value backwards to the point where the values revert to zero is the alignment shown by BestFit. This alignment is the best segment of similarity between the two sequences.

For nucleic acids, the default scoring matrix has a *match* value of 1.0 for each identical symbol comparison and -0.90 for each non-identical comparison (not considering nucleotide ambiguity symbols for this example). The *quality* score for a nucleic acid alignment can, therefore, be determined using the following equation:

```
Quality = 1.0 x TotalMatches +  -0.90 x TotalMismatches
                  - (GapCreationPenalty x GapNumber)
          - (GapExtensionPenalty x TotalLengthOfGaps)
```

The *quality* score for a protein alignment is calculated in a similar manner. However, while the default nucleic acid scoring matrix has a single value for all non-identical comparisons, the default protein scoring matrix has different values for the various non-identical amino acid comparisons. The *quality* score for a protein alignment can therefore be determined using the following equation (where $Total_{AA}$ is the total number of A-A (Ala-Ala) matches in the alignment, $CmpVal_{AA}$ is the value for an A-A comparison in the scoring matrix, $Total_{AB}$ is the total number of A-B (Ala-Asx) matches in the alignment, $CmpVal_{AB}$ is the value for an A-B comparison in the scoring matrix, ...) :

$$
\begin{aligned}
Quality = {} & CmpVal_{AA} \times Total_{AA} \\
& + CmpVal_{AB} \times Total_{AB} \\
& - CmpVal_{AC} \times Total_{AC} \\
& \vdots \\
& - CmpVal_{zz} \times Total_{zz} \\
& - (GapCreationPenalty \times GapNumber) \\
& - (GapExtensionPenalty \times TotalLengthOfGaps)
\end{aligned}
$$

For a more complete discussion of scoring matrices, see the Data Files manual.

CONSIDERATIONS

BestFit Always Finds Something

BestFit always finds an alignment for any two sequences you compare -- even if there is no significant similarity between them! You must evaluate the results critically to decide if the segment shown is not just a random region of relative similarity.

The Segments Shown Obscure Alternative Segments

BestFit only shows one segment of similarity, so if there are several, all but one is obscured. You can approach this problem with graphic matrix analysis (see the Compare and DotPlot programs). Alternatively, you can run BestFit on ranges outside the ranges of similarity found in earlier runs to bring other segments out of the shadow of the best segment.

The Best Fit is Only One Member of a Family

Like all fast gapping algorithms, the alignment displayed is a member of the family of best alignments. This family may have other members of equal quality, but will not have any member with a higher quality. The family is usually significantly different for different choices of gap creation and gap extension penalties. See the CONSIDERATIONS topic in the entry for the Gap program in the **Program Manual** to learn more about how to assign gap creation and gap extension penalties.

The Surface of Comparison

The magnitude of the computer's job is proportional to the area of the surface of comparison. That area is determined by the product of the lengths of the two sequences compared. BestFit can evaluate a surface of up to 3.5 million elements. This surface would be large enough to compare two sequences approximately 1,870-symbols long, or one sequence 200-symbols long with another sequence 17,500-symbols long. When you have much longer sequences that are known to align well, you can use the command-line option `-LIMit` to use the surface more efficiently.

The Public Scoring Matrix for Nucleic Acid Comparisons is Very Stringent

The scoring matrix swgapdna.cmp penalizes mismatches -0.9 so the segments found may be very brief. This penalty means that the alignment cannot be extended by three bases to pick one extra match. The scoring matrix used by Smith and Waterman, when local alignments were first described, used -0.333 for the mismatch penalty. You can use Fetch to copy randomdna.cmp and rename it swgapdna.cmp to use these values, or use nwsgapdna.cmp, which has no mismatch penalty at all.

Rapid Alignment

When possible, BestFit tries to find the optimal alignment very quickly. If this rapid alignment is not unambiguously optimal, BestFit automatically realigns the sequences to calculate the optimal alignment. When this occurs, the monitor of alignment progress on your terminal screen (`Aligning...`) is displayed twice for a single alignment.

ALIGNING LONG SEQUENCES

This program can align very long sequences if you know roughly where the alignment of interest begins. Run the program with the command line option `-LIMit`. Then set the starting coordinates for each sequence near the point where the alignment of interest begins and set gap shift limits on each sequence. The program then aligns the sequences from your starting point such that the sequences do not get out of phase by more than the gap shift limits you have set. If you started both sequences at

base number one and set the gap shift limit for sequence one to 100 and for sequence two to 50, then base 350 in sequence one could not be gapped to any base outside of the range from 300 to 450 on sequence two.

If you omit `-LIMit` on the command line, the program automatically sets gap shift limits if they are needed to allow the alignment of long sequences to proceed. In this case, the program limits the total length of gaps that can be inserted into each sequence and calculates the best alignment within this incomplete, or *limited*, surface of comparison. The program then performs a calculation to determine whether the alignment could possibly be improved if there were no restriction on the total length of gaps in each sequence. If the program cannot rule out this possibility, it displays the message `*** Alignment is not guaranteed to be optimal ***`. Because the criteria used in the calculation for guaranteeing an optimal alignment are very stringent, a limited alignment often may be optimal even if this message is displayed. In any event, the program continues to completion.

EVALUATING ALIGNMENT SIGNIFICANCE

This program can help you evaluate the significance of the alignment, using a simple statistical method, with the `-RANdomizations` command line option. The second sequence is repeatedly shuffled, maintaining its length and composition, and then realigned to the first sequence. The average alignment score, plus or minus the standard deviation, of all randomized alignments is reported in the output file. You can compare this average *quality* score to the quality score of the actual alignment to help evaluate the significance of the alignment. The number of randomizations can be specified along with the `-RANdomizations` command line qualifier; the default is 10.

The score of each randomized alignment is reported to the screen. You can use `<Ctrl>C` to interrupt the randomizations and output the results from those randomized alignments that have been completed.

By ignoring the statistical properties of biological sequences, this simple Monte Carlo statistical method may give misleading results. Please see Lipman, D.J, Wilbur, W.J., Smith, T.F., and Waterman, M.S. (Nucl. Acids Res. 12; 215-226 (1984)) for a discussion of the statistical significance of nucleic acid similarities.

ALIGNMENT METRICS

BestFit and Gap display four figures of merit for alignments: Quality, Ratio, Identity, and Similarity.

The Quality (described above) is the metric maximized in order to align the sequences. Ratio is the quality divided by the number of bases in the shorter segment. Percent Identity is the percent of the symbols that actually match. Percent Similarity is the percent of the symbols that are similar. Symbols that are across from gaps are ignored. A similarity is scored when the scoring matrix value for a pair of symbols is greater than or equal to 0.50, the *similarity threshold*. This threshold is also used by the display procedure to decide when to put a ':' (colon) between two aligned symbols. You can reset it from the command line with the second optional parameter of `-PAIr`. For instance, the expression `-PAIr=1.0,0.5` would set the similarity threshold to 0.5.

The similarity and identity metrics are not optimized by alignment programs so they should not be used to compare alignments.

PEPTIDE SEQUENCES

If your input sequences are peptide sequences, this program uses a scoring matrix with matches scored as 1.5 and mismatches scored according to the evolutionary distance between the amino acids as measured by Dayhoff and normalized by Gribskov (Gribskov and Burgess Nucl. Acids Res. 14(16); 6745-6763 (1986)).

RESTRICTIONS

Input sequences may not be more than 30,000-symbols long. This program cannot evaluate a surface of comparison larger than 5.5 million elements. A 200 x 27,500 comparison is possible, as well as a 2,300 x 2,300 comparison. See the ALIGNING LONG SEQUENCES topic for help in aligning long sequences that would normally exceed the maximum surface of comparison. You can also ask your system manager to increase the maximum surface of comparison if your system has enough virtual memory.

SEQUENCE TYPE

The function of BestFit depends on whether your input sequence(s) are protein or nucleotide. Normally the type of a sequence is determined by the presence of either `Type: N` or `Type: P` on the last line of the text heading just above the sequence itself. If your sequence(s) are not the correct type, turn to Appendix VI for information on how to change or set the type of a sequence.

COMMAND-LINE SUMMARY

All parameters for this program may be put on the command line. Use the option `-CHEck` to see the summary below and to have a chance to add things to the command line before the program executes. In the summary below, the capitalized letters in the qualifier names are the letters that you *must* type in order to use the parameter. Square brackets ([and]) enclose qualifiers or parameter values that are optional. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the **User's Guide**.

```
Minimal Syntax: % bestfit [-INfile1=]gamma.seq  [-INfile2=]alu.seq -Default

Prompted Parameters:

-BEGin1=1 -BEGin2=1       beginning of each sequence
-END1=500 -END2=207       end of each sequence
-NOREV1     -NOREV2       strand of each sequence
-GAPweight=5.0            gap creation penalty     (3.0 is protein default)
-LENgthweight=0.3         gap extension penalty    (0.1 is protein default)
[-OUTfile1=]gamma.pair    output file for alignment

Local Data Files: -DATa=swgapdna.cmp  scoring matrix for nucleic acids
                  -DATa=swgappep.cmp  scoring matrix for peptides

Optional Parameters:

-OUTfile2=gamma.gap       new sequence file for sequence 1 with gaps added
-OUTfile3=alu.gap          "    "       "   "      "    2   "   "     "
-LIMit1=499 -LIMit2=206 limit the surface of comparison
-RANdomizations[=10]      determine average score from 10 randomized
                             alignments
-PAIr=1.0,0.5,0.1         thresholds for displaying '|', ':', and '.'
-WIDth=50                 the number of sequence symbols per line
-PAGe=60                  adds a line with a form feed every 60 lines
-NOBIGGaps                suppresses abbreviation of large gaps with '.'s
-HIGhroad                 makes the top alignment for your parameters
-LOWroad                  makes the bottom alignment for your parameters
-NOSUMmary                suppresses the screen summary
```

ACKNOWLEDGEMENTS

Gap and BestFit were originally written for Version 1.0 by Paul Haeberli from a careful reading of the Needleman and Wunsch (J. Mol. Biol. 48; 443-453 (1970)) and the Smith and Waterman (Adv. Appl. Math. 2; 482-489 (1981)) papers.

Limited alignments were designed by Paul Haeberli and added to the Package for Version 3.0. They were united into a single program by Philip Delaquess for Version 4.0. Default gap penalties for protein alignments were modified according to the suggestions of Rechid, Vingron and Argos (CABIOS 5; 107-113 (1989)).

LOCAL DATA FILES

The files described below supply auxiliary data to this program. The program automatically reads them from a public data directory unless you either 1) have a data file with exactly the same name in your current working directory; or 2) name a file on the command line with an expression like `-DATa1=myfile.dat`. For more information see Chapter 4, Using Data Files in the **User's Guide**.

If the first sequence you name is a nucleic acid, BestFit uses the scoring matrix in the public file swgapdna.cmp. (SW stands for Smith and Waterman.) If the first sequence you name is a peptide sequence, BestFit reads swgappep.cmp instead. The presence of these files in your current working directory causes BestFit to read your version instead. (See the **Data Files** manual for more information about scoring matrices.)

OPTIONAL PARAMETERS

The parameters and switches listed below can be set from the command line. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the **User's Guide**.

`-LIMit1=20 and -LIMit2=20`

let you set *gap shift limits* for each sequence. When you already know of a long similarity between two sequences you can "zip" them together using this mode. The beginning coordinates for each sequence must be near the beginning of the alignment you want to see. The alignment continues so that gaps inserted do not require the sequences to get out of step by more than the gap shift limits. You can align very long sequences rapidly. The surface of comparison is still limited to 3.5 million. The size of a comparison can be predicted by multiplying the average length of the two sequences by the sum of the two shift limits.

If you add `-LIMit` to the command line without any qualifier value, the program prompts you to enter gap shift limits for each sequence.

`-RANdomizations=10`

reports the average alignment score and standard deviation from 10 randomized alignments in which the second sequence is repeatedly shuffled, maintaining the length and composition of the original sequence, and then aligned to the first sequence. You can use the optional parameter to set the number of randomized alignment to some number other than 10.

`-OUTfile2=seqname1.gap -OUTfile3=seqname2.gap`

This program can write three different output files. The first displays the alignment of sequence one with sequence two. The second is a new sequence file for sequence one, possibly expanded by gaps to make it align with sequence two. The third, like the second, is a new sequence file for sequence two, possibly expanded by gaps to make it align with sequence one. The program writes only the first file unless there are output file options on the command line. If there are any output files named on the command line, *only* those output files are written. If you add

-OUT to the command line without any qualifying filename, then the program will write the second and third output files after prompting you for their names.

Aligned sequences (in sequence files) can be displayed with GapShow. Their similarity can be displayed with PlotSimilarity.

-PAIr=1.0,0.5,0.1

The paired output file from this program displays sequence similarity by printing one of three characters between similar sequence symbols: a pipe character(|), a colon (:), or a period (.). Normally a pipe character is put between symbols that are the same, a colon is put between symbols whose comparison value is greater than or equal to 0.50, and a period is put between symbols whose comparison value is greater than or equal to 0.10. You can change these *match display thresholds* from the command line. The three parameters for -PAIr are the display thresholds for the pipe character, colon, and period. The match display criterion for a pipe character changes from symbolic identity (the default) to the quantitative threshold you have set in the first parameter. A pipe character will no longer be inserted between identical symbols unless their comparison values are greater than or equal to this threshold. If you still want a pipe character to connect identical symbols, use x instead of a number as the first parameter. (See the Data Files manual for more information about scoring matrices.)

-PAGe=64

When you print the output from this program, it may cross from one page to another in a frustrating way -- especially when you print on individual sheets. This option adds form feeds to the output file in order to try to keep clusters of related information together. You can set the number of lines per page by supplying a number after the -PAGe qualifier.

-WIDth=50

puts 50 sequence symbols on each line of the output file. You can set the width to anything from 10 to 150 symbols.

-NOBIGGaps

suppresses large gap abbreviations, showing all the sequence characters across from large gaps. Usually, gaps that extend one sequence by more than one complete line of output are abbreviated with three dots arranged in a vertical line.

-LOWroad and -HIGhroad

The insertion of gaps is, in many cases, arbitrary, and equally optimal alignments can be generated by inserting gaps differently. When equally optimal alignments are possible, this program can insert the gaps differently if you select either the -LOWroad or the -HIGhroad options. Here are examples for the alignment of GACCAT with GACAT with different parameters.

```
For:        Match = 1.0        MisMatch = -0.9
       Gap weight = 1.0   Length Weight =  0.0

LowRoad:   1 GACCAT 6
             .  |||       Quality = 4.0
           1 GA.CAT 5

HighRoad:  1 GACCAT 6
             ||| ||       Quality = 4.0
           1 GAC.AT 5
```

```
For:        Match = 1.0        MisMatch =  0.0
      Gap weight = 3.0   Length Weight =  0.0

HighRoad:  1 GACCAT 6
             |||              Quality = 3.0
           1 GACAT. 5

LowRoad:   1 GACCAT 6
              |||             Quality = 3.0
           1 .GACAT 5
```

Essentially the *low road* shifts all of the arbitrary gaps in sequence two to the left and all of the arbitrary gaps in sequence one to the right. The *high road* does exactly the opposite. When neither *high road* nor *low road* is selected, the program tries not to insert a gap whenever that is possible and uses the high road alternative for all collisions.

-SUMmary

writes a summary of the program's work to the screen when you've used the -Default qualifier to suppress all program interaction. A summary typically displays at the end of a program run interactively. You can suppress the summary for a program run interactively with -NOSUMmary.

Use this qualifier also to include a summary of the program's work in the log file for a program run in batch.

Oncogene (1997) 15, 1671-1680
 0950-9232/97 $12.00

Characterization of changes in gene expression associated with malignant transformation by the NF-κB family member, v-Rel

Oleksi Petrenko, Irene Ischenko and Paula J Enrietto

Department of Microbiology, State University of New York at Stony Brook, Stony Brook, New York, 11794, USA

In this study, alterations in gene expression patterns have been examined in v-Rel-transformed avian bone marrow cells. Using a conditional v-Rel estrogen receptor chimera (v-RelER) which transforms cells in an estrogen-dependent manner, we constructed subtraction cDNA libraries from v-RelER-transformed bone marrow cells. Several different sequences were identified whose expression was altered upon hormone activation of v-RelER. These include two genes related to the MIP-1 chemokine family (*mip-1β* and a *tca3* homologue), a cell surface antigen *sca-2* and the transcription factor *nfkb1*. The expression of each gene was assayed in a number of wild-type and mutant v-Rel-expressing fibroblast and hematopoietic cells. All v-Rel-transformed hematopoietic cells tested express high levels of *nfkb1* and *sca-2*. In fibroblasts, wild-type v-Rel induced expression of *mip-1β* and *nfkb1*, while nontransforming mutants of v-Rel failed to do so, suggesting a role for these two genes in v-Rel mediated transformation. Finally, these genes are expressed at high levels in cells overexpressing wild-type and truncated forms of c-Rel, implying that v-Rel transforms, in part, by induction of c-Rel target genes.

Keywords: v-Rel; NF-kappa B; oncogene; transformation

Introduction

The Rel/NF-κB transcription factor family includes proteins structurally related through an amino terminal region, the Rel Homology Domain. This region is largely responsible for several properties of the proteins, including homo- and heterodimer formation, DNA binding, and interaction with a family of inhibitor proteins. All members of the Rel/NF-κB family have been implicated in the regulation of transcription by virtue of their interaction with the κB enhancer, a potent cis-regulatory sequence present in many inducible cellular and viral genes. Active NF-κB transcription complexes are homo- and heterodimers consisting of one or two members of the Rel/NF-κB family, which includes NF-κB1, NF-κB2, RelA, RelB and c-Rel (Grilli *et al.*, 1993; Siebenlist *et al.*, 1994). The activity of NF-κB complexes is regulated by a second family of proteins, the IκB family (Beg and Baldwin, 1993; Verma *et al.*, 1995).

The target genes for Rel/NF-κB regulation are numerous, and are, for the most part, involved in cellular growth and immunoregulatory processes (Grilli *et al.*, 1993; Baeuerle and Baltimore, 1996). Alterations in several members of the family have been associated with hematopoietic malignancies. Thus, the genes encoding c-Rel, RelA, NF-κB1, NF-κB2 and Bcl-3 are located at sites of genomic rearrangements in certain human cancers (Lu *et al.*, 1991; Neri *et al.*, 1991; Liptay *et al.*, 1992; Ohno *et al.*, 1993). Targeted disruption of c-*rel*, *relb*, and *nfkb1* leads to multiple functional defects in the immune system (Kontgen *et al.*, 1995; Sha *et al.*, 1995; Weih *et al.*, 1995), while the disruption of *rela* and *ikba* results in embryonic or neonatal lethality (Beg *et al.*, 1995; Klement *et al.*, 1996).

Given the role that Rel/NF-κB proteins play in normal and oncogenic processes, it is critical to understand the mechanism by which altered Rel expression generates the leukemic phenotype. v-Rel, a mutated homologue of c-Rel, remains the most carefully studied Rel/NF-κB family member with respect to its oncogenic potential. Isolated as the oncogene within the avian retrovirus, REV-T, v-Rel induces a rapidly fatal hematopoietic malignancy in birds and transforms fibroblasts, splenic cultures and bone marrow hematopoietic progenitor cells *in vitro* (Bose, 1992). Previous studies revealed that v-Rel forms protein complexes with other Rel/NF-κB proteins and binds to NF-κB motifs *in vitro* (Gilmore *et al.*, 1996). While early works suggested that v-Rel acts as a dominant negative mutant of c-Rel (Gilmore, 1990; Bose, 1992), recent data indicate that transformation by v-Rel results from its capacity to positively or negatively alter expression of genes important for hematopoietic cell growth and differentiation (Baeuerle and Baltimore, 1996; Gilmore *et al.*, 1996).

To more fully understand the mechanism of v-Rel transformation, we investigated the transcriptional changes induced in hematopoietic cells transformed by a conditional form of v-Rel. Fusion of v-Rel to the hormone-binding domain of human estrogen receptor (v-RelER, Boehmelt *et al.*, 1992) resulted in the creation of a chimeric protein whose biological and biochemical properties were inducible and indistinguishable from wild-type v-Rel. Utilizing subtractive cDNA libraries derived from v-RelER-transformed hematopoietic cells grown in the presence or absence of estrogen, we identified several different sequences whose expression was altered upon hormone activation of v-RelER. These include two genes related to the MIP-1 chemokine family (*mip-1β* and a *tca3* homologue), a cell surface antigen *sca-2*, and the transcription factor *nfkb1*. The expression of each of these genes was assayed in a number of wild-type and mutant v-Rel-expressing fibroblast and hematopoietic cells. All v-Rel-transformed hematopoietic cells tested expressed high levels of *nfkb1* and *sca-2*. In fibroblasts, wild-type

Correspondence: O Petrenko
Received 14 February 1997; revised 4 June 1997; accepted 4 June 1997

v-Rel induced expression of *mip-1β* and *nfkb1*, while nontransforming mutants of v-Rel failed to do so, suggesting a role for these two genes in v-Rel-mediated transformation. Finally, these genes are expressed at high levels in cells overexpressing wild-type and truncated forms of c-Rel, implying that v-Rel transforms, in part, by induction of c-Rel target genes.

Results

Analysis of subtraction cDNA libraries

To understand the changes in gene expression that accompany transformation by v-Rel, we sought to identify genes whose levels change upon hormone activation of v-RelER. This was accomplished by the construction of subtractive cDNA libraries from estrogen-induced v-RelER cells and from the cells withdrawn from estrogen (see Materials and methods). Hybridization and sequence analysis of the clones isolated from the subtraction libraries permitted the identification of twelve different cDNAs. The determined sequences of clones 59 (IκBα), 256 (NF-κB1) and 393 (CAP-23) were identical to the corresponding sequences of chicken cDNAs. The resolved sequence of clone 220 exhibited 80% homology with the Drosophila actin-related protein. Other identified genes revealed high level homology with the corresponding mammalian sequences. Table 1 summarizes the clones identified and the size of the corresponding mRNA transcripts.

One group of genes, present at 3–10-fold higher levels in estrogen-stimulated v-RelER cells, included the two putative cytokines, Macrophage Inflammatory protein 1β (MIP-1β, clone 4), and a homologue of TCA 3 (hereafter referred to as cTCA, chicken T cell activation protein, clone 391); clone 44, ornitine decarboxylase antizyme (ODC-Az), a key regulator of ornitine decarboxylase which is constitutively activated in various transformed cells (Auvinen *et al.*, 1992); clone 59, IκBα; clone 71, a member of the STAT family of signal transducers and activators of transcription, Stat1; clone 80, a chicken homologue of the mammalian Stem Cell Antigen-2 (Sca-2); clone 214, a regulatory subunit of protein phosphatase 2A (PP2A); clone 220, a homologue of the maternally loaded Drosophila embryo actin-related protein (ARP) that plays a role in early embryogenesis (Frankel *et al.*, 1994); NF-κB1, the Rel/NF-κB family member; CAP-23, a cortical cytoskeleton-associated protein found in developing neural tissues (Widmer and Caroni, 1990).

Of the clones that were reproducibly more abundant in the v-RelER cells withdrawn from estrogen, we were able to identify two different sequences. These genes encode translation initiation factor-2α (eIF-2α), which promotes binding of initiator tRNAs to 40S ribosomal subunits (Ernst *et al.*, 1987), and nucleosome assembly protein-1 (NAP-1) involved in transcription factor binding and nucleosome displacement (Walter *et al.*, 1995).

Table 1 cDNA clones isolated from the subtraction libraries

Clone number	*Expression in v-RelER cells (+Er)*	*(−Er)*	*Size mRNA (kb)*	*Sequence homology*
4	+ +	−	1.0	MIP-1β (mouse)
44	+ +	+	1.6	ODC-Az (rat)
59	+ +	+	3.0	IκBα (chicken)
71	+ + +	+	>4.0	Stat1 (human)
80	+ + +	+	1.3	Sca-2 (mouse)
214	+ +	+	2.4/4.0	PP2A (rabbit)
220	+ +	+	2.0	ARP (Drosophila)
256	+ + +	+	3.9	NF-κB1 (chicken)
391	+ + +	−	0.8	TCA3 (mouse)
393	+ + +	+	1.1	CAP-23 (chicken)
31	+	+ +	1.8/4.0	eIF-2α (human)
513	+	+ +	2.6	NAP1 (mouse)

Expression levels were classified into four groups as follows: + + +, 0.2–0.4% of total mRNA, strong expression; + +, 0.04–0.1%, moderate expression; +, 0.02% and less, weak expression; −, no detectable expression

Expression of differentially regulated genes in v-RelER cells

The differential expression of identified genes in v-RelER cells was confirmed by Northern blot analysis of mRNAs prepared from cells grown in the presence or absence of estrogen. Because *nfkb1* was isolated in this screen as a potential v-Rel-regulated gene, the expression of other Rel/NF-κB family members was also examined, including c-*rel*, *rela*, *relb* and *nfkb2*. In addition, we examined the expression of a gene which decreases during hematopoietic cell differentiation, c-*myb* (Graf, 1992). One class of genes, including *ctca*, *mip-1β*, *stat1*, *nfkb1* and *nfkb2*, was expressed in v-RelER cells grown in the presence of estrogen, and significantly downregulated within 1 day of estrogen withdrawal (see Figure 1). The expression of a second class of genes, typified by *cap-23*, *sca-2*, *ikba* and c-*rel*, decreased less dramatically upon estrogen withdrawal, suggesting that additional factors are involved in their regulation. Two other Rel family members, *rela* and *relb*, were expressed below the level of detection in v-RelER cells (data not shown). The expression of three genes, *eif-2α*, *nap-1*, and c-*myb*, increased upon withdrawal of v-RelER cells from estrogen. c-Myb, a sequence-specific DNA-binding protein, is thought to regulate genes whose expression is incompatible with cellular differentiation as reflected in its expression pattern and biological properties (Graf, 1992). Surprisingly, we found that c-*myb* mRNA increased upon estrogen withdrawal, conditions under which v-RelER cells begin to differentiate into dendritic and neutrophilic cells (Boehmelt *et al.*, 1995).

Transcriptional induction of rel-*regulated genes*

To further investigate the possibility that genes isolated in this screen are under transcriptional control by v-Rel, the v-RelER cells maintained without estrogen for 3 days were estrogen-induced for various periods of time. Subsequently, expression of the corresponding mRNAs was analysed. While not dividing, v-RelER bone marrow cells withdrawn from estrogen were metabolically active, as judged by their ability to revert phenotypically upon the readdition of estrogen (Boehmelt *et al.*, 1992) and by expression of *β-actin*, *GAPDH*, *vimentin*, *MHC class II* mRNAs at levels equivalent to hormone induced cells (data not shown).

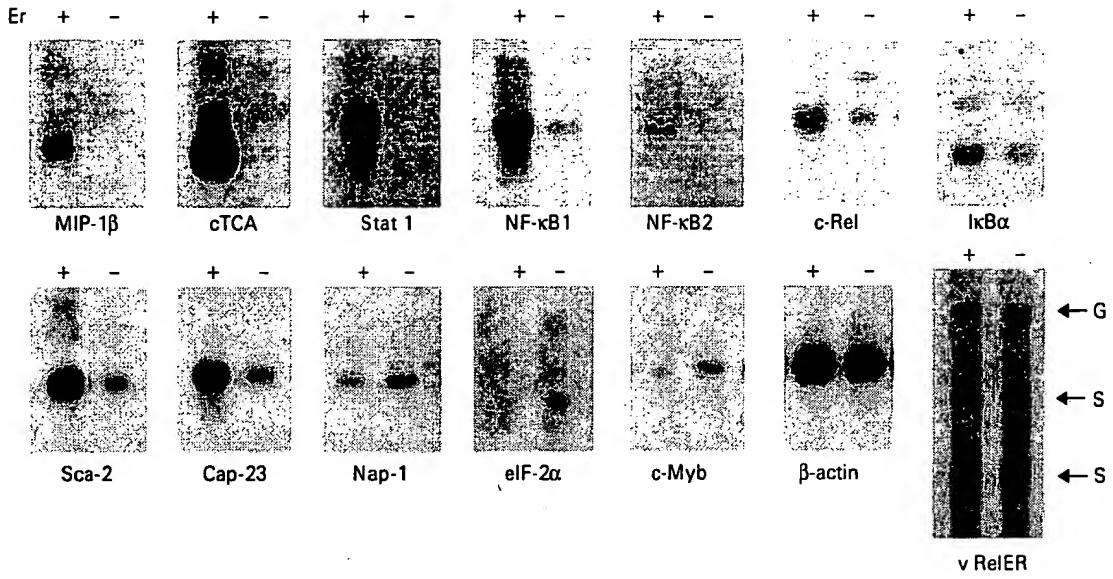


Figure 1 Expression of differentially regulated genes in v-RelER bone marrow cells grown in the presence of estrogen (Er+), or without estrogen for 1 day (Er−). Genomic (G) RCASv-RelER and spliced (S) v-*relER* mRNAs are indicated

As expected from the results described above, *mip-1β*, *ctca* and *nfkb1* expression was restored within 4 h of estrogen addition (see Figure 2). Similar results were obtained with *nfkb2*. These data are consistent with the time course of activation observed in our previous work on the localization and DNA binding properties of estrogen-stimulated v-RelER (Boehmelt *et al.*, 1992). A more complicated picture was observed for the induction of *sca-2* mRNA. Although expression of *sca-2* was dependent on estrogen activated v-RelER, its levels were less responsive to estrogen activation, demonstrating a delayed kinetics of activation. Other examined genes, including *arp*, *cap-23*, *pp2A* and *stat1*, were not rapidly induced by hormone, indicating that they were not directly regulated by v-RelER (data not shown).

Expression patterns of rel*-regulated genes*

To exclude the possibility that the expression of the isolated genes reflected a general transformed state, not v-Rel transformation, the expression patterns were analysed in hematopoietic cells transformed by various oncogenes. In addition to genes rapidly induced by hormone, *ikba* and three other genes, c-*rel*, c-*myb* and *stat-1*, encoding transcription factors whose expression was altered in v-RelER cells were included in this analysis (see Table 2). Two genes downregulated in estrogen activated v-RelER cells, *eif-2a* and *nap-1*, were found at similar levels in most lines tested, including v-Rel-transformed NPB4 cells. *Mip-1β*, c-*rel* and c-*myb*, were expressed at elevated levels in v-Ski-transformed cells. Other examined genes, including *ctca*, *nfkb1*, *nfkb2*, *stat1*, *sca-2* and *ikba*, were found at higher levels in NPB4 cells, pointing to their potential role in v-Rel-mediated transformation. No evidence for lineage-specific expression of these genes could be observed. Instead, v-Ski-transformed precursors for the erythroid and myeloid lineages displayed elevated levels of several of the same genes as v-Rel-transformed lymphoid cells (see Table 2), which may be due to the elevated levels of c-Rel in v-Ski cells.

Previous studies established that the target cell for v-Rel transformation depends on the virus complex used for infection. It has been reported that v-Rel transforms mature IgM-positive B lymphocytes, an IgM-negative cell within the T cell or myeloid lineages, or a cell expressing markers of the myeloid and B cell lineages (Barth and Humphries, 1988; Barth *et al.*, 1990; Morrison *et al.*, 1991). To provide evidence for a correlation between expression of these genes and v-Rel transformation, we assayed gene expression in a variety of v-Rel-transformed cells derived from different hematopoietic lineages. As can be seen from Table 3, which summarizes this analysis, some variations in the expression levels were found, particularly with *mip-1β* and *ctca*. Thus, *ctca* was expressed in 8 out of 10 cell lines at variable levels, while *mip-1β* was found in most lines at low levels. Two other genes, c-*myb* and *nfkb2*, were expressed at low levels in most tested cells (data not shown). In contrast, *nfkb1*, *stat1*, *ikba* and *sca-2* were found at elevated levels in almost all cells tested, suggesting that overexpression of these genes is characteristic of v-Rel-induced phenotype.

These experiments also allowed us to compare the ability of v-Rel and overexpressed or truncated c-Rel to induce altered gene expression. Two hematopoietic cell lines, B-1 and 189/5, were examined which contain both wild-type and carboxy terminal truncated forms of c-Rel (Hrdlickova *et al.*, 1994). Each cell line expresses high levels of all genes tested (see Table 3). Because B-1 cells contain high levels of both wild-type and truncated form of c-Rel, it is not clear if transcriptional activation and transformation in these cells result from wild-type or mutant c-Rel activity. In contrast, 189/5 cells contain low levels of wild-type c-Rel and high levels of carboxy terminal truncated form of c-Rel, suggesting that altered gene expression results from the truncated version of c-Rel.

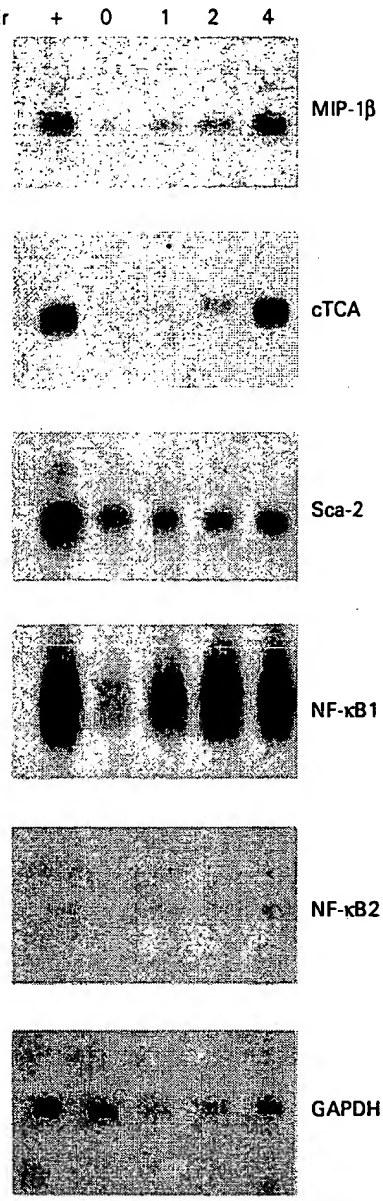


Figure 2 Time course of gene induction by estrogen-activated v-RelER. The v-RelER-transformed bone marrow cells grown in the presence of estrogen (+) were withdrawn from hormone for 3 days (0). The cells were subsequently estrogen induced for the indicated periods of time over the period of 4 h and cellular RNAs were probed with the corresponding cDNAs indicated on the right

Induction of gene expression by carboxy terminal mutants of v-rel

Examination of the biological properties of v-Rel mutants has shown that transformation required sequences within the amino terminal Rel Homology Domain. Small carboxy terminal deletions have a marginal effect on the transforming properties of the protein (Sarkar and Gilmore, 1993; Smardova *et al.*, 1995). In our previous studies, deletions approximately 100 bp in length were made throughout v-*rel* and mutant proteins were analysed in CEF and BM cells (see Materials and methods). All deletion mutants within the *rel*-homology domain (*dl2-dl8*) were transformation-defective, as none of them were able to bind to DNA. However, mutants which lie outside the *rel*-homology domain (*dl1*, *dl12-17*) retained the ability to bind to DNA, activate transcription of cellular genes, and transform fibroblasts and BM cells to different degrees (Morrison *et al.*, 1992; Smardova *et al.*, 1995). Taking advantage of these mutants, we correlated gene expression with the transforming ability of v-*rel*. Bone marrow from 4–10-day old chicks was infected with wild-type v-*rel* or carboxy terminal v-*rel* deletion mutants (*dl12,13,15-17*) and subsequently maintained in liquid culture. All cultures expanded to approximately 10^9 cells within several weeks. Analysis of gene expression showed that *mip-1β*, *nfkb1* and *stat1* were expressed in all cells tested, though some variations in expression levels were observed (see Figure 3). In contrast, most v-*rel* mutations resulted in low levels of *ctca* and *sca-2* expression. In our previous work, decreased levels of two other Rel-regulated genes, MHC class I and HMG 14b, were observed in bone marrow cells infected with the carboxy terminal v-*rel* deletion mutants. Though it is possible that each deletion mutant transformed a different cell type, the phenotypic analysis of these cells suggested that this is not the case (Smardova *et al.*, 1995).

Expression of rel-*regulated genes in CEFs*

Wild-type v-Rel and estrogen-activated v-RelER confer characteristic growth properties and morphology on primary CEFs (Morrison *et al.*, 1991; Boehmelt *et al.*, 1992). Because nontransforming mutants of v-*rel* do not promote growth of bone

Table 2 Expression of the identified clones in transformed hematopoietic cell lines

	NPB4 (lymphoid)	*DT95 (lymphoid)*	*HP50 (lymphoid)*	*v-ski BM (myeloid, erythroid)*	*BM2 (myeloid)*	*HD11 (myeloid)*	*HD3 (erythroid)*
Clone number	v-*rel*	*ND**	*ND**	v-*ski*	v-*myb*	v-*myc*	v-*erbA*/*ts*-v-*erbB*
MIP-1β	+	–	–	+ +	–	+	–
cTCA	+ + +	–	–	–	–	–	–
NF-κB1	+ +	+	+	+	+	+	+
NF-κB2	+	–	–	–	–	–	–
Sca-2	+ +	+	–	+	–	–	+
IκBα	+ +	–	+	+ +	+	+	+
c-Rel	+	+	+	+ +	–	–	+
Stat1	+ +	–	–	+ +	–	–	–
c-Myb	–	–	–	+ +	–	–	–
eIF-2α	+	NA	+	+	+	+	NA
NAP1	+ +	NA	+	+ +	+	+	NA

ND*, lymphoid cell lines derived from ALV induced tumors; NA, not analysed. For explanation of expression levels see footnotes to Table 1

Table 3 Expression of genes in v-Rel transformed cells

Cell line	Virus	Tissue and lineage derivation	v-Rel 1	c-Rel 2	NFκB1 3	IκBα 4	Stat1 5	MIP-1β 6	cTCA 7	Sca-2 8
RCAS-1	RCASv-Rel	BM, myeloid	+ +	+	+ +	+ +	+ +	+	+ +	+ + +
RCAS-2	RCASv-Rel	BM, myeloid	+ +	+	+	+ +	+ + +	+	+ +	+ +
NPB4	REV-T/REV-A	BM, lymphoid	+ + +	+	+ +	+ +	+ +	+	+ + +	+ +
BM1	REV-T/REV-A	BM, lymphoid	+ +	+	+ +	+ +	+ +	+	+ +	+ +
tu-1	REV-T/REV-A	embryo liver, nd	+ + +	+	+ +	+ +	+ +	+	+ +	+ +
tu-2	REV-T/REV-A	embryo liver, nd	+ + +	+	+ +	+ +	+ +	+	+ +	+ +
SS-1	REV-T/CSV	spleen, B-lymphoid	+ + +	+	+ +	+ +	+ + +	+	–	+ +
123/6T	REV-TW/REV-A	spleen, nonB/nonT	+ + +	+	+ +	+ +	+ +	–	–	+ +
123/6	REV-TW/CSV	spleen, nonB/nonT	+ + +	+	+ +	+ +	+ +	–	+	+ +
160/2	REV-TW/CSV	spleen, T-lymphoid	+ +	+	+ +	+ +	+ +	+	+	+ +
189/5	REV-C/CSV	spleen, T-lymphoid	–	+ + +	+ +	+ +	+ + +	+ +	+ + +	+ +
B-1	REV-C/CSV	spleen, T-lymphoid	–	+ + +	+ + +	+ +	+ + +	+ +	+ + +	+ + +

BM, bone marrow; nd, not defined. For explanation of expression levels see footnotes to Table 1

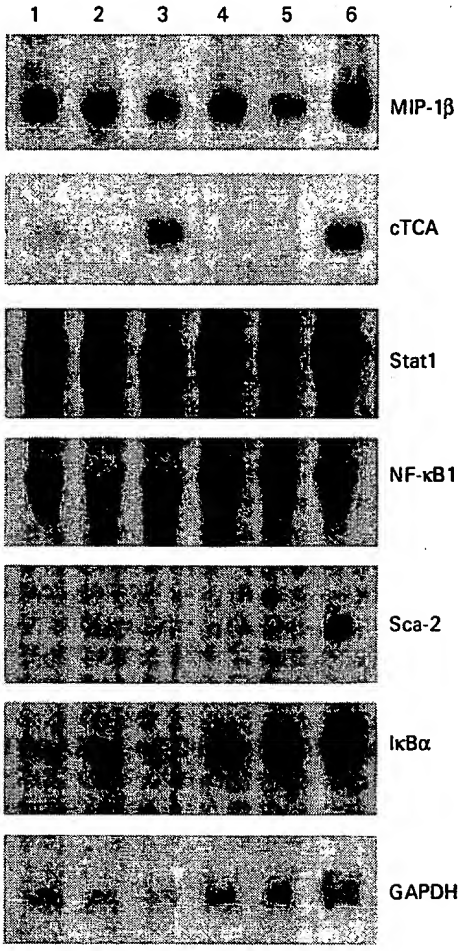


Figure 3 Northern blot analysis of RNA isolated from bone marrow cells infected with v-*rel* carboxy terminal deletion mutants *dl12* (lane 1), *dl13* (lane 2), *dl15* (lane 3), *dl16* (lane 4), *dl17* (lane 5), or wild-type v-*rel* (lane 6). An equivalent filter was probed with GAPDH to control for equal loading

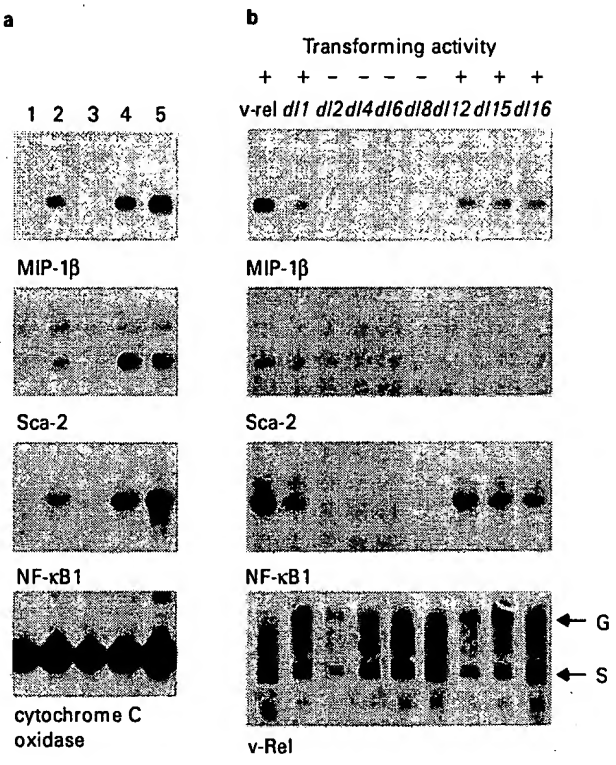


Figure 4 Analysis of *mip-1β*, *sca-2* and *nfκb1* expression in primary CEFs. (a) Induction of gene expression in CEFs transfected with RCAS (lane 1), RCASv-Rel (lane 2), RCASv-RelER (cells grown in the absence of estrogen, lane 3), RCASv-RelER (cells grown in the presence of estrogen, lane 4), or RCASc-Rel (lane 5). A filter containing each of the RNAs was probed with the mitochondrial cytochrome oxidase cDNA to control for equal loading. (b) Induction of gene expression in CEFs infected with wild-type v-*rel* or v-*rel* deletion mutants. Geneomic (G) RCASv-Rel and spliced (S) v-*rel* mRNAs are indicated

marrow cells, we examined gene expression in avian fibroblasts infected with mutant forms of v-*rel*. Initially, the inducibility of *mip-1β*, *sca-2* and *nfkb1* was tested in v-RelER fibroblasts maintained with or without estrogen. As can be seen in Figure 4, all three mRNAs were expressed in v-RelER cells grown in the presence of estrogen. Fibroblasts overexpressing v-Rel and wild-type c-Rel were also analysed and found to express each of the tested genes (Figure 4a). In contrast, *ctca*, *nfkb2* and *stat1* were not found in fibroblasts presumably because hematopoietic cell-specific factors are required for their expression (data not shown).

Next, gene expression was examined in fibroblasts expressing transforming (v-*rel*, *dl-1*, *dl-12*, *dl-15* and *dl-16*) or nontransforming mutants of v-*rel* (*dl-2*, *dl-4*, *dl-6* and *dl-8*). As can be seen in Figure 4b, the expression of *mip-1β* and *nfkb1* was evident only in cells producing wild-type v-*rel*. *dl-1* mutant which lacks

amino terminal viral *env* sequences, or carboxy terminal transforming v-*rel* mutants (*dl12-dl16*). Non-transforming mutants within the Rel Homology Domain (*dl2-dl8*) did not induce expression of these genes. Interestingly, while the expression of *sca-2* was inducible in fibroblasts, it was only found in wild-type v-*rel*-transformed cells. All v-*rel* mutants failed to induce expression of this gene, reflecting the complexity of the protein structure and interactions required for full function of v-Rel.

Characterization of viral-transduced fibroblasts

To further investigate the role of MIP-1β and NF-κB1 in cellular growth control, the effect of overexpression of these genes was studied in chicken fibroblasts. The corresponding cDNAs were subcloned into the retroviral vector pCRNCM, downstream of the CMV promoter. These constructs, together with c-*rel* and c-*kit* cDNAs used as controls, were introduced into chicken embryo fibroblasts. The recombinant viruses were subsequently produced by infecting the cells with the helper transformation-defective virus tdB77. CEFs infected with the recombinant viruses were selected in Geneticin and characterized on the RNA or protein levels. To coexpress the corresponding *rel* proteins, the cells were superinfected with either RCASv-Rel or RCASc-Rel (see Figure 5).

Analysis of the growth phenotypes revealed that the fibroblasts overexpressing *nfkb1* maintained the growth properties characteristic of control pCRNCM-transduced CEFs. These cells reproducibly grew for 20–25 passages as did control cells; after that they became vacuolated and died (see Table 4). In contrast, overexpression of *mip-1β* extended life the span of CEFs. In addition, the *mip-1β* fibroblasts efficiently proliferated in low serum and displayed the ability to anchorage-independent growth in soft agar characteristic of transformed cells. Coexpression of v-*rel* in these cells further promoted their ability to form colonies in soft agar.

As reported earlier, overexpression of c-*rel* morphologically transforms CEFs (Abbadie *et al.*, 1993; Kralova *et al.*, 1994). Although c-*rel* fibroblasts do not form colonies in soft agar, they do display characteristic disruption of cellular cytoskeleton and a remarkable extension of life span. Interestingly, CEFs overexpressing both c-*rel* and *mip-1β* retained the major growth properties of the c-*rel* fibroblasts. In addition, they efficiently formed colonies in soft agar, revealing a fully transformed phenotype (see Table 4). In sharp contrast, coexpression of c-*rel* and v-*rel* in fibroblasts resulted in decreased cell growth, consistent with our previous study which showed that v-Rel and c-Rel interfere with the DNA binding properties of each other (Hodgson and Enrietto, 1995).

Discussion

Most evidence available to date indicates that transformation by v-Rel results from its capacity to form homodimers, enter the nucleus, and bind DNA. While early studies suggested that v-Rel could repress transcription, presumably by interfering with c-Rel

Table 4 Growth properties of viral-transduced fibroblasts

Transfected cDNAs[a]	*Virus*	*Growth rate (h/cell division)*	*Colony formation in soft agar*	*Life span (passages)*[b]
pCRNCM	tdB77	48	–	20–25
c-Rel	tdB77	56	–	30–35
NF-κB1	tdB77	46	–	20–25
MIP-1β	tdB77	42	8%	35–40
c-Kit	tdB77	38	3%	nd
RCASv-Rel*	RCASv-Rel	39	–	50–60
c-Rel	RCASv-Rel	68	–	10–15
NF-κB1	RCASv-Rel	46	–	25–30
MIP-1β	RCASv-Rel	42	15%	40–45
c-Kit	RCASv-Rel	42	6%	nd
RCASc-Rel*	RCASc-Rel	50	–	40–50
NF-κB1	RCASc-Rel	44	–	25–30
MIP-1β	RCASc-Rel	42	20%	40–45
c-Kit	RCASc-Rel	39	5%	nd

[a]All cDNAs were subcloned into the pCRNCM retroviral vector, except where indicated (*). To coexpress the corresponding *rel* proteins, transfected CEFs were superinfected with either RCASv-Rel or RCASc-Rel. [b]Cells that grew for 20–25 passages were in culture approximately 2 months. Cells that grew for 50–60 passages were in culture for over 6 months

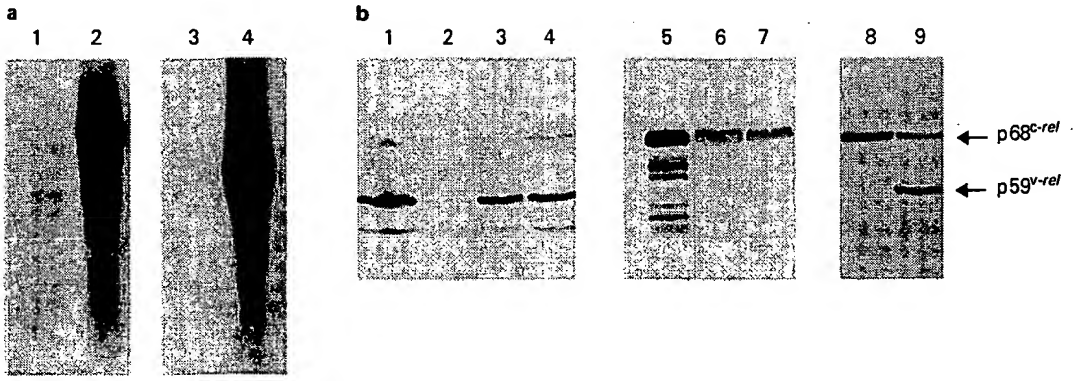


Figure 5 Analysis of gene expression in viral-transduced fibroblasts. (**a**) Northern blot analysis of RNA isolated from normal CEFs (lanes 1 and 3), NF-κB1 CEFs (lane 2), and MIP-1β CEFs (lane 4). Lanes 1 and 2 were hybridized with the *nfkb1* probe; lanes 3 and 4 were probed with *mip-1β*. (**b**) Immunoblot analysis of Rel expression in RCASv-Rel CEFs (lane 1); NF-κB1 CEFs and MIP-1β CEFs superinfected with RCASv-Rel (lanes 3 and 4, respectively); RCASc-Rel CEFs (lane 5); NF-κB1 CEFs and MIP-1β CEFs superinfected with RCASc-Rel (lanes 6 and 7, respectively); c-Rel CEFs (lane 8); c-Rel CEFs superinfected with RCASv-Rel (lane 9). Expression in nontransfected CEFs (lane 2) is shown as a control

activity (Gilmore, 1990; Bose, 1992), more recent works point to a role for v-Rel in the transcriptional activation of a number of endogenous genes important for hematopoietic cell growth and differentiation (Baeuerle and Baltimore, 1996; Gilmore *et al.*, 1996).

These studies and our previous work promoted us to develop a more complete picture of the genes whose expression might contribute to the v-Rel-induced leukemic phenotype. Taking advantage of the conditional Rel variant, subtraction cDNA libraries were constructed from v-RelER-transformed bone marrow cells grown in the presence or absence of estrogen. Examination of the expression patterns of the isolated genes suggested that several of them might be directly induced by v-Rel upon transformation.

Two Rel/NF-κB family members were upregulated in estrogen-stimulated v-RelER cells, *nfkb1* and *nfkb2*. Activation of these two genes in lymphoid cells expressing either v-Rel or c-Rel was reported earlier (Hrdlickova *et al.*, 1995). Recent evidence suggests that heterodimers of NF-κB2/p52 with v-Rel can play a role in v-Rel-mediated transformation. Thus, heterodimers containing p52 and nontransforming mutants of v-Rel, that cannot form homodimers, can transform chicken spleen cells (White *et al.*, 1996). In addition, coexpression of aberrant versions of p52 increases the oncogenicity of c-Rel proteins with carboxy terminal deletions (Gilmore *et al.*, 1996). While our results demonstrate that the *nfkb2* expression levels remained low in most examined cell lines, the expression of *nfkb1* was high in all examined v-Rel-transformed hematopoietic cells and correlated with v-Rel transformation in mutant v-*rel*-infected cells. A number of studies indicate that NF-κB1/p105 plays a role of a cytoplasmic retention molecule for the Rel/NF-κB proteins (Beg and Baldwin, 1993; Grilli *et al.*, 1993). On the other hand, the processing of p105 may lead to different amounts of p50 homo- and heterodimers in the cell. While in v-Rel-transgenic mice, NF-κB1 is not required for transformation (Carasco *et al.*, 1996), previous studies have associated elevated levels of p50 with several lymphoid and non-lymphoid malignancies (Mukhopadhyay *et al.*, 1995; Bargou *et al.*, 1996). NF-κB1/p50 may contribute to oncogenesis by forming heterodimeric complexes with v-Rel, c-Rel, and p52, which function as potent transcription regulators (Siebenlist *et al.*, 1994; Baldwin, 1996). The cells expressing constitutively active NF-κB, such as B cells or HTLV-I-transformed T cells, contain primarily c-Rel/p50 heterodimers in their nuclei, while in transformed cells, v-Rel is predominantly complexed with p50 and IκBα (Liou *et al.*, 1994; Miyamoto *et al.*, 1994; Gilmore *et al.*, 1996). The prevalence of constitutively active v-Rel/p50 in these cells may explain the upregulation of NF-κB target genes such as cytokines and their receptors, MHC, and cell adhesion molecules. Thus, constitutive expression of *nfkb1* induced by v-Rel may result in 'constitutive' activation of the transformed cell.

Two isolated cDNAs, clones 4 and 391, revealed sequence homology with the family of chemoattractant cytokines. Clone 4 is a chicken homologue of the mammalian macrophage inflammatory protein-1 beta, while clone 391 is the homologue of the mouse T cell activation protein 3, a cytokine structurally related to MIP-1β and IL-8 (Burd *et al.*, 1987). Examination of the expression patterns of both cytokines revealed that most v-Rel-transformed hematopoietic cells, regardless of lineage, express *mip-1β*. Its expression also correlated with transformation when v-*rel* deletion mutants were examined in bone marrow cells and fibroblasts. In contrast, *ctca* was not expressed in v-Rel-transformed fibroblasts. In addition, its expression failed to correlate with transformation in bone marrow cells, suggesting that it is not required for the maintenance of v-Rel-mediated transformation. In summary, our data argue for a role for *mip-1β* in transformation by v-Rel and a direct role for v-Rel in its transcriptional regulation.

Previous studies showed that MIP-1β is a potent chemoattractant for monocytes and specific subpopulations of lymphocytes (Taub *et al.*, 1995; Lloyd *et al.*, 1996). Specifically, MIP-1β stimulates T cell proliferation and induces actin polymerization and profound cytoskeletal changes in T cells within seconds of exposure (Adams *et al.*, 1994). It also exhibits growth regulatory properties for hematopoietic stem and progenitor cells, costimulating myelopoiesis and antagonizing the growth inhibitory activity of MIP-1α (Graham *et al.*, 1990). Our data indicate that overexpression of *mip-1β* in avian fibroblasts induces cellular transformation as measured by growth in soft agar and extended life span of the cells. However, in most examined v-Rel-transformed hematopoietic cell lines, *mip-1β* was expressed at low to moderate levels. Therefore, its relevance to the transforming process of v-Rel is yet to be demonstrated.

The role of the chicken Sca-2 homologue in transformation remains unclear, in part because its biological and biochemical properties are not well defined. *Sca-2* expression failed to correlate with v-Rel transformation when v-*rel* deletion mutants were examined in fibroblast and hematopoietic cells. In addition, its Rel-inducibility was not clear in v-RelER cells. However, we found that infection of chicken lymphoid or erythroid cell lines with RCASv-Rel induces the expression of *sca-2* (data not shown). These results imply that the regulation of *sca-2* expression is complex and may involve factors other than v-Rel. Mammalian Sca-2 is normally present on a subset of immature thymic blasts and bone marrow cells that repopulate the thymus (Sprangrude *et al.*, 1989). Recent data suggest a role for Sca-2 in T cell activation and protection from TCR-mediated apoptosis (Saitoh *et al.*, 1995; Noda *et al.*, 1996). Because *sca-2* was found at high levels in all v-Rel-transformed hematopoietic cells we cannot exclude the possibility that it plays a role in the generation of v-Rel-induced phenotype.

Several v-Rel target genes have been described so far, including HMG-14b and MHC class I in bone marrow cells (Boehmelt *et al.*, 1992; Walker and Enrietto, 1996); NF-κB1 and IκBα in fibroblasts and lymphoid lines (Hrdlickova *et al.*, 1995; Schatzle *et al.*, 1995); IL-2R, DM-GRASP, p75, MHC class I and II, identified through introduction of v-*rel* into *myc*-transformed lymphoid cell lines (Hrdlickova *et al.*, 1994; Zhang *et al.*, 1995; Zhang and Humphries, 1996). In contrast, MHC class I and IκBα were not upregulated in tumor cells from v-Rel-transgenic mice (Carrasco *et al.*, 1996). Our previous work showed

that MHC class II was not regulated by v-Rel in transformed bone marrow cells (Walker and Enrietto, 1996). While the subset of Rel-regulated genes may vary in different cell types, the data presented here indicate that v-Rel acts by stimulating the expression of a number of genes, some of which are regulated by c-Rel. These alterations may be the result of the direct action of v-Rel in regulation of genes such as *mip-1β*, inappropriate expression of which may potentiate growth of the transformed cell. Alternatively, v-Rel may act indirectly, altering the expression of genes through the upregulation of other transcription factors, such as NF-κB1. Most likely, development of the v-Rel-induced transformed phenotype is the sum of all these changes, each of which may have profound biological consequences on the target cell. For this reason, we are currently assessing the contribution of each of these genes to the transformed phenotype.

In the course of this work, we also examined the effect of overexpression of wild-type c-Rel on the genes identified in this screen. *Mip-1β*, *sca-2* and *nfkb1* were all upregulated in c-Rel-overexpressing CEF, suggesting that genes normally controlled by c-Rel are targets for v-Rel regulation. Since both v-Rel and c-Rel overexpression result in morphological transformation of CEF, perhaps expression of this set of the genes is required to mediate fibroblast transformation. Two c-Rel-transformed spleen cell lines were also examined in this study and found to have elevated levels of *nfkb1*, *sca-2*, *stat1*, *mip-1β* and *ctca*. Because these cell lines also contain truncated forms of c-Rel (Hrdlickova *et al.*, 1994), it is not clear if transcriptional activation and transformation in both of them result from wild-type or mutant c-Rel activity. We previously reported that overexpression of wild-type c-Rel in bone marrow cells, which appear to be granulocytic in origin (unpublished observation), leads to cell death (Abbadie *et al.*, 1993). For this reason, it will be important to determine if the genes identified in this study are upregulated in the granulocytic bone marrow target prior to the onset of cell death.

Materials and methods

Cells and tissue culture

CEFs expressing different viruses were prepared using the calcium-phosphate/DNA transfection method (Chen and Okayama, 1987). Generation of v-Rel and v-RelER transformed bone marrow (BM) cells was previously described (Morrison *et al.*, 1991; Boehmelt *et al.*, 1992). The origin of other cell lines is as follows: BM2, an immature myeloid cell line transformed by v-Myb (Moscovici *et al.*, 1977); HD11, a myeloid cell line transformed by v-Myc (Leutz *et al.*, 1984); HP50 and DT95, lymphoid cell lines derived from ALV induced tumors (Kabrun *et al.*, 1990; Hrdlickova *et al.*, 1994); v-Ski transformed precursors for the erythroid and myeloid lineages (Larsen *et al.*, 1993); HD3 erythroblasts transformed by v-ErbA/ts-v-ErbB (Schmidt *et al.*, 1986); 189/5 and B-1, T-lymphoid lines transformed by c-Rel (Hrdlickova *et al.*, 1994); BM1, tu-1, tu-2 (P Enrietto, unpublished), NPB4 (Beug *et al.*, 1981), SS-1, 123/6, 123/6T, and 160/2 (Hrdlickova *et al.*, 1994) are v-Rel transformed cell lines.

v-rel deletion mutants

Generation and characterization of v-*rel* deletion mutants have been described previously (Morrison *et al.*, 1992; Smardova *et al.*, 1995). Briefly, deletions approximately 100 bp in length were made by oligonucleotide-directed mutagenesis throughout v-*rel*. Each mutant was cloned into the RCAS vector for analysis in CEFs and BM cells. All deletion mutants between nucleotides 37 and 798 in v-*rel* (*dl2–dl8* in this study) were transformation defective. Mutants that lie outside this region retained the ability to transform fibroblasts and BM cells to different degrees (*dl1*, *dl12–dl17*).

Characterization of growth properties

The cDNAs encoding *mip-1β*, *nfkb1*, c-*rel* or c-*kit*, were subcloned into the pCRNCM retroviral vector (Metz *et al.*, 1991), downstream of the CMV promoter. The retroviral constructs were transfected into CEFs and selected in Geneticin. Recombinant viruses were produced by infecting the cells with tdB77 transformation-defective helper virus. The viruses were harvested after 4 days of cultivation. CEFs infected with the recombinant viruses were selected in Geneticin and characterized for the corresponding RNA or protein levels. Proliferation of fibroblasts was measured in DMEM supplemented with 1% bovine serum and 0.2% chicken serum by direct counting of trypsinized cells using Coulter counter. Proliferation related to tumorigenesis was examined in soft agar medium composed of DMEM, 6% fetal calf serum, 2% chicken serum, 1 × MEM vitamin solution, penicillin (100 u/ml), streptomycin (100 μg/ml), and nystatin (100 u/ml). Trypsinized cells were mixed with a 0.25% agar-medium solution to a concentration of 2×10^5 cells/ml and 3 ml of each cell mixture was added to 60 mm Petri dishes prepared with an initial 3 ml underlay of 0.7% agar. Cells were grown for 14–16 days in a humidified incubator at 37°C under 5% CO_2. Proliferation was evaluated by the microscopic counting of individual colonies (average of four experiments).

Construction and screening of cDNA libraries

The poly(A)$^+$RNA prepared from v-RelER-transformed BM cells was copied into cDNAs and cloned into the *Eco*RI or *Not*I site of lambda ZAPII vector (Stratagene). Two subtraction cDNA libraries were prepared: one enriched for v-RelER-induced sequences (library I), and the other enriched for the sequences downregulated by v-RelER (library II). *In vitro* excision of single-stranded Bluescript phagemids containing cDNA inserts, biotinylation and subtractions were performed as described (Schweinfest *et al.*, 1990). Following subtractive hybridization, the phagemids were transfected into *E. coli* XL1-Blue cells (Stratagene). Approximately 600–700 bacterial colonies were obtained as a result of each subtraction. Bacterial clones were tested by differential Northern hybridization with cellular RNAs and subjected to sequence analysis. DNA sequencing of double-stranded plasmid templates was performed using Sequenase kit (USB). At least 400 bp of each clone were sequenced to determine homologies. Nucleotide sequences were analysed using Blast database search programs (NCBI).

Northern blot hybridization

RNAs prepared using acid guanidinium thiocyanate-phenol extraction procedure (Chomczynski and Sacchi, 1987) were separated on formaldehyde-agarose gels and blotted onto the Hybond-N nylon membranes (Amersham). Hybridization probes were derived by *pcr*

amplification of the cloned full-length cDNAs encoding MIP-1β, cTCA, Sca-2, CAP-23 and PP2A; partial cDNAs for ARP, eIF-2α, NAP-1, ODC-Az and Stat-1, including the corresponding coding and 3'-untranslated regions; a 2.5 kb long NF-κB1 cDNA (clone 256) that lacks the 5' *rel*-homology region. Other hybridization probes included v-*rel* in the plasmid pBSrelCS (Morrison *et al.*, 1991); c-*rel* in the plasmid pBSc-Rel (Abaddie *et al.*, 1993); c-*myb* in the plasmid pneoCCC (kindly provided by Dr J Lipsick); MHC class I cDNA clone B-F12aF10 (Boehmelt *et al.*, 1992); MHC class II cDNA clone BLBbII (kindly provided by Dr Ch Auffray); NF-κB1, NF-κB2 and RelB cDNAs were kindly provided by Dr TD Gilmore; IκBα cDNA was kindly provided by Dr IM Verma.

Representative Northern blots were quantified using Ultroscan laser densitometer (LKB). The levels were normalized to the density of hybridization of the serial dilutions of the corresponding cDNA plasmids run in parallel with the RNA samples. Expression levels of c-Rel, v-Rel, NF-κB1 and Stat-1 in transformed hematopoietic cell lines were verified by immunoblot analysis using the corresponding antibodies.

Western blot analysis

Protein extracts for Western blot analysis were prepared as described (Morrison *et al.*, 1991). The antibodies used were SB66 Rel-specific polyclonal antibody (Boehmelt *et al.*, 1992), a polyclonal anti-avian NF-κB1 antibody (kindly provided by Dr HR Bose), and anti-Stat1 mAb (Transduction Laboratories).

Accession numbers

The sequences described have the following GenBank accession numbers: clone 4, L34553; clone 80, L34554; clone 391, L34552.

Acknowledgements

We would like to thank Dr MJ Hayman for critical reading of the manuscript. We also thank Kathleen Donnely and Linda Whittaker for technical support. This work was supported by a grant from the Council for Tobacco Research (#3196). Further support came from the National Institutes of Health (CA51792).

References

Abbadie C, Kabrun N, Boulai F, Smardova J, Stehelin D, Vandenbunder B and Enrietto PJ. (1993). *Cell*, **75**, 899–912.

Adams DH, Harvath L, Bottaro DP, Interrante R, Catalano G, Tanaka Y, Strain A, Hubster SG and Shaw S. (1994). *Proc. Natl. Acad. Sci. USA*, **91**, 7144–7148.

Auvinen M, Paasinen A, Anderson LC and Holtta E. (1992). *Nature*, **360**, 355–358.

Baeuerle P and Baltimore D. (1996). *Cell*, **87**, 13–20.

Baldwin AS. (1996). *Annu. Rev. Immunol.*, **14**, 649–681.

Bargou RC, Leng C, Krappman D, Emmerich F, Mapara MY, Bommert K, Royer HD, Scheidereit C and Dorken B. (1996). *Blood*, **87**, 4340–4347.

Barth CF, Ewert DL, Olson WC and Humphries EH. (1990). *J. Virol.*, **64**, 6054–6062.

Barth CF and Humphries EH. (1988). *J. Exp. Med.*, **167**, 89–108.

Beg AA and Baldwin AS. (1993). *Genes Dev.*, **7**, 2064–2070.

Beg AA, Sha WC, Bronson RT, Ghosh S and Baltimore D. (1995). *Nature*, **376**, 167–170.

Beug H, Muller H, Doederlein G and Graf T. (1981). *Virology*, **115**, 295–309.

Boehmelt G, Walker A, Kabrun N, Mellitzer G, Beug H, Zenke M and Enrietto PJ. (1992). *EMBO J.*, **11**, 4641–4652.

Boehmelt G, Madruga J, Dorfler P, Briegel K, Schwarz H, Enrietto PJ and Zenke M. (1995). *Cell*, **80**, 341–352.

Bose HR. (1992). *Biochim. Biophys. Acta.*, **1114**, 1–17.

Burd R, Freeman GJ, Wilson SD, Berman M, DeKruyff R and Dorf ME. (1987). *J. Immunol.*, **139**, 3126–3131.

Carrasco D, Rizzo CA, Dorfman K and Bravo R. (1996). *EMBO J.*, **15**, 3640–3650.

Chen C and Okayama H. (1987). *Mol. Cell. Biol.*, **7**, 2745–2752.

Chomczynski P and Sacchi N. (1987). *Anal. Biochem.*, **162**, 156–159.

Ernst H, Duncan RF and Hershey JW. (1987). *J. Biol. Chem.*, **262**, 1206–1212.

Frankel S, Heintzelman MB, Artavanis-Tsakonas S and Mooseker MS. (1994). *J. Mol. Biol.*, **235**, 1351–1356.

Gilmore TD. (1990). *Cell*, **62**, 841–843.

Gilmore TD, Koedood M, Piffat KA and White DW. (1996). *Oncogene*, **13**, 1367–1378.

Graf T. (1992). *Curr. Opin. Genet. Dev.*, **2**, 249–255.

Graham GJ, Wright EG, Hewick R, Wolpe SD, Wilkie NM, Donaldson D, Lorimore S and Pragnell IB. (1990). *Nature*, **344**, 442–444.

Grilli M, Chiu JJS and Lenardo MJ. (1993). *Int. Rev. Cytol.*, **143**, 1–62.

Hodgson J and Enrietto PJ. (1995). *J. Virol.*, **69**, 1971–1979.

Hrdlickova R, Nehyba J and Humphries E. (1994). *J. Virol.*, **68**, 2371–2382.

Hrdlickova R, Nehyba J, Roy A, Humphries EH and Bose HR. (1995). *J. Virol.*, **69**, 403–413.

Kabrun N, Bumstead N, Hayman M and Enrietto PJ. (1990). *Mol. Cell. Biol.*, **10**, 4788–4794.

Klement JF, Rice NR, Car BD, Abbondanzo SJ, Powers GD, Bhatt PH, Chen CH, Rosen CA and Stewart CL. (1996). *Mol. Cell. Biol.*, **16**, 2341–2349.

Kontgen F, Grumont RJ, Strasser A, Metcalf D, Li R, Tarlinton D and Gerondakis S. (1995). *Genes Dev.*, **9**, 1965–1977.

Kralova J, Schatzle JD, Bargmann W and Bose HR. (1994). *J. Virol.*, **68**, 2073–2083.

Larsen J, Mayer S, Steinlein P, Beug H and Hayman M. (1993). *Oncogene*, **8**, 3221–3228.

Leutz A, Beug H and Graf T. (1984). *EMBO J.*, **3**, 3191–3114.

Liou C, Sha WC, Scott ML and Baltimore D. (1994). *Mol. Cell. Biol.*, **14**, 5349–5359.

Liptay S, Schmid RM, Perkins ND, Metzer P, Alther MR, McPherson JD, Wasmuth JJ and Nabel GJ. (1992). *Genomics*, **13**, 287–292.

Lloyd AR, Oppenheim JJ, Kelvin DJ and Taub DD. (1996). *J. Immunol.*, **156**, 932–938.

Lu D, Thompson JD, Gorski GK, Rice NR, Meyer MG and Yunis J. (1991). *Oncogene*, **6**, 1235–1241.

Metz T, Graf T and Leutz A. (1991). *EMBO J.*, **10**, 837–844.

Miyamoto S, Chiao PJ and Verma IM. (1994). *Mol. Cell. Biol.*, **14**, 3276–3282.

Morrison LE, Boehmelt G, Beug H and Enrietto PJ. (1991). *Oncogene*, **6**, 1657–1665.

Morrison LE, Boehmelt G and Enrietto PJ. (1992). *Oncogene*, **7**, 1137–1147.

Moscovici C, Moscovici MG, Jiminez H, Lai MM, Hayman MJ and Vogt PK. (1977). *Cell*, **11**, 95–103.

Mukhopadhyay T, Roth JA and Maxwell SA. (1995). *Oncogene*, **11**, 999–1003.

Neri A, Chang CC, Lombardi L, Salina M, Corrandi P, Maiolo AT, Chaganti RSK and Dalla-Fevera R. (1991). *Cell*, **67**, 1075–1087.

Noda S, Kosugi A, Saito S, Narumiya S and Hamaoka T. (1996). *J. Exp. Med.*, **183**, 2355–2360.

Ohno H, Yambumoto K, Fukuhara S and McKeithan TW. (1993). *Leukemia*, **7**, 2057–2063.

Saitoh S, Kosugi A, Noda S, Yamamoto N, Ogata M, Minami Y, Miyake K and Hamaoka T. (1995). *J. Immunol.*, **155**, 5574–5581.

Sarkar S and Gilmore TD. (1993). *Oncogene*, **8**, 2245–2252.

Schatzle JD, Kralova J and Bose HR. (1995). *J. Virol.*, **69**, 5383–5390.

Schmidt JA, Marshall J, Hayman MJ, Doederlei G and Beug H. (1986). *Leuk. Res.*, **10**, 257–272.

Schweinfest C, Henderson KW, Gu JR, Kottaridis S, Besbeas S, Panotopoulou E and Papas T. (1990). *Genet. Anal. Techn. Appl.*, **7**, 64–70.

Sha WC, Liou HC, Tuomanan EI and Baltimore D. (1995). *Cell*, **80**, 321–330.

Siebenlist U, Franzoso G and Brown K. (1994). *Annu. Rev. Cell Biol.*, **10**, 405–455.

Smardova J, Walker A, Morrison LE, Kabrun N and Enrietto PJ. (1995). *Oncogene*, **10**, 2017–2026.

Sprangrude GJ, Klein J, Heimfeld S, Aihara Y and Weissman I. (1989). *J. Immun.*, **142**, 425–430.

Taub DD, Sayers TJ, Carter CR and Ortaldo JR. (1995). *J. Immunol.*, **155**, 3877–3888.

Verma IM, Stevenson JK, Schwarz EM, van Antverp D and Miyamoto S. (1995). *Genes Dev.*, **9**, 2723–2735.

Walker A and Enrietto PJ. (1996). *Oncogene*, **12**, 2515–2525.

Walter PP, Owen-Hughes TA, Cote J and Workman JL. (1995). *Mol. Cell. Biol.*, **15**, 6178–6187.

Weih F, Carrasco D, Durham SK, Barton DS, Rizzo CA, Ryceck RP, Lira SA and Bravo R. (1995). *Cell*, **80**, 331–340.

White DW, Pitoc GA and Gilmore TD. (1996). *Mol. Cell. Biol.*, **16**, 1169–1178.

Widmer F and Caroni P. (1990). *J. Cell Biol.*, **111**, 3035–3047.

Zhang G, Slaughter C and Humphries E. (1995). *Mol. Cell. Biol.*, **15**, 1806–1816.

Zhang G and Humphries EH. (1996). *Oncogene*, **12**, 1153–1157.

Nucl. Med. Biol. Vol.14, No.4, pp.277-280, 1987
Int. J. Radiat. Appl. Instrum. Part B
Printed in Great Britain

EXHIBIT "D"

0883-2897/87 $3.00+0.00
Pergamon Journals Ltd

SYNTHETIC PEPTIDES AND MONOCLONAL ANTIBODIES IN IMMUNOASSAYS

PROTEIN STRUCTURE AND ANTIGENICITY

Marc H.V. VAN REGENMORTEL

Department of Immunochemistry,
Institute of Molecular and Cellular Biology, Strasbourg, France

ABSTRACT

The antigenicity of a protein resides in a series of mutually overlapping surface patches known as epitopes which make contact with the combining sites of antibody molecules. Epitopes are usually localized by demonstrating the presence of antigenic cross-reactivity between a protein and some peptide fragments. Since structural features of proteins such as the accessibility, hydrophilicity and mobility of segments of the polypeptide chain have been correlated with the location of epitopes, it is possible to predict from the primary structure which linear peptides are likely to correspond to epitopes of the protein.

The antigenicity of a protein refers to its capacity to bind specifically to the functional binding sites or paratopes of certain immunoglobulin molecules. Paratopes are made up of six highly accessible loops of hypervariable sequence that interact to a greater or lesser extent, with the surface of the antigen. That portion of the antigen that comes into contact with the paratope of the antibody constitutes an antigenic determinant or epitope of the antigen. In the same way that the antibody nature of an immunoglobulin is identified only after its complementary antigen has been recognized, the epitope nature of a cluster of amino acids in a protein can be established only by using an immunoglobulin as a detecting device. An epitope is thus a relational entity which can be defined only in a functional and operational sense through the availability of complementary paratopes. The question to be addressed here is : to what extent do structural features of a protein correlate with the location of its epitopes ?

It has been customary to divide epitopes into a number of conceptual categories such as sequential and conformational epitopes that are not easily distinguished experimentally. At the present time, it is common to distinguish between continuous and discontinuous epitopes (Benjamin et al., 1984 ; Berzofsky, 1985 ; Van Regenmortel, 1986). Continuous epitopes consist of amino acid residues in direct peptide linkage while discontinuous epitopes are defined as a cluster of residues that are not contiguous in sequence but are juxtaposed at the protein surface by the folding of the polypeptide chain.

A common approach for identifying epitopes in a protein consists in measuring the ability of natural or synthetic fragments of the molecule to react with antibodies raised against the complete molecule. Any linear peptide of 5-10 residues that is found to react is labelled a continuous epitope. However, such a label should not be taken to imply that the linear fragment accurately mimics the complete structure of the corresponding epitope in the native protein. In most cases, such a linear peptide is likely to represent only part of a larger discontinuous epitope of the protein ; antibodies directed to a discontinuous epitope may indeed react, albeit weakly, with subregions of the epitope made up of a few residues in linear sequence. Although most of our knowledge of protein antigenicity is based on the identification of

continuous epitopes, it should be emphasized that they represent only incomplete and adulterated versions of the original epitopes existing in the native protein. It is widely believed that the majority of epitopes are of the discontinuous type (Benjamin et al., 1984 ; Berzofsky, 1985 ; Barlow et al., 1986), although until now only a single epitope of this type has been fully delineated in lysozyme (Amit et al., 1986). In this case, 16 residues of lysozyme (segments 18-27 and 116-129 of the sequence) were found by X-ray crystallography of antigen-antibody complexes to be in contact with 17 residues of the paratope belonging to all six complementarity-determining regions. The two complementary surfaces showed extensive interpenetration and mutual interdigitation.

In general, information on discontinuous epitopes is extremely patchy, since only two or three amino acids can usually be shown to be contact residues of the epitope ; this information is obtained by showing that related proteins presenting substitutions at these positions are discriminated by a particular monoclonal antibody (Benjamin et al., 1984 ; Van Regenmortel, 1984).

Many practical applications of immunological research are based on the exploitation of antigenic cross-reactions, made possible for instance by raising antibodies against synthetic peptides able to cross-react with the complete protein molecule (Lerner, 1984 ; Walter, 1986).

The most systematic way to look for continuous epitopes in proteins is to synthesize all possible overlapping hexa-, hepta- or octapeptides of the protein, and then to measure their capacity to bind to antiprotein antibodies. This is readily achieved by a method developed by Geysen et al. (1984) in which the peptides are synthesized on a linear polymer of polyacrylic acid and tested immunologically while still bound to the solid phase. The contribution, to the antibody-binding interaction, of individual amino acids within a synthetic peptide can then be examined by systematically substituting each residue by the other 19 possible amino-acids (Rodda et al., 1986). In this way, it can be shown that a certain number of residues of continuous epitopes are essential for antigenicity (i.e. they cannot be replaced by any other amino acid) while other residues can be replaced by all common amino acids without affecting peptide binding by the antibody (Getzoff et al., 1987). When this method was applied to myoglobin (Rodda et al., 1986), a new epitope undetected in earlier work (Atassi, 1984) was identified in residues 48-55. However, several myoglobin epitopes identified previously by other immunochemical techniques (Atassi, 1984) were not revealed by the method of Geysen. These discrepancies demonstrate the operational nature of any definition of antigenicity resulting from the fact that the type of probe as well as the particular immunoassay used greatly influence the result. In a recent study of the histone H2A (Muller et al., 1986), it was shown, for instance, that the antigenic activity of several synthetic peptides depended on whether they were tested as free peptides in solution, conjugated to a carrier, or adsorbed to a plastic solid phase. Such variations are probably due to variations in peptide accessibility and conformation in the different assays.

In recent years, there have been many attempts to correlate the position of continuous epitopes in proteins with certain features of their primary, secondary or tertiary structures (Parker et al., 1986). For instance, plots of hydrophilicity (Hopp, 1986) along the peptide chain are often used to identify linear stretches of residues that are exposed to the solvent and therefore likely to be antigenic.

The surface location of many chain termini in proteins (Thornton & Sibanda, 1983) probably explains why the surface terminal residues are so often implicated in protein epitopes. However, it is also possible that the antigenicity of chain termini could be due to their high relative mobility compared to other more constrained sections of the polypeptide chain (Westhof et al., 1984).

It has been shown that segments of highest local mobility (with an amplitude of only a few Å) correlate with highly accessible regions at the surface of the protein (e.g. loops or reverse turns) as well as with the positions of continuous epitopes of a length of 5-10 residues (Westhof et al., 1984 ; Tainer et al., 1985). In a recent study, a number of continuous epitopes

of myohemerythrin were identified by measuring the ability of synthetic hexapeptides to bind to myohemerythrin antibodies (Geysen et al., 1987). The study included all 113 possible hexapeptides encompassing the 118 residues of the protein. The relative degree of antigenicity of the peptides was assessed by the number of rabbits immunized with the protein that responded to each peptide as well as by mean antiserum titre. The five peaks in the mobility curve were found to correspond to regions of higher than average antigenicity. It seems that antibodies preferentially tend to recognize short peptides when these correspond to mobile segments of a native protein. It should be emphasized that the magnitude of the motions found in segmental mobility is small (1-2 Å) and that contrary to some claims (Novotny & Haber, 1986), the energetic cost for binding is therefore not necessarily prohibitive. Movements of a few Å within the epitope may have a beneficial effect on the critical positioning of residues and this could result in an induced fit and increased binding affinity (Edmundson & Ely, 1986). Such a dynamic view of immunological complementarity is in line with the wide-spread recognition that the functional activity of proteins is often linked to dynamic conformational changes. However, some protein chemists and crystallographers reject such a dynamic interpretation and argue that the location of epitopes simply correlates with the most exposed regions of the protein surface (Fanning et al., 1986 ; Novotny & Haber, 1986). Proponents of this viewpoint believe that static surface accessibility is sufficient to explain the location of epitopes in proteins and they consider accessibility and mobility as two mutually exclusive explanatory categories. There is, however, good evidence that static accessibility is not always a necessary and sufficient condition for antigenicity (Wilson et al., 1984 ; Van Regenmortel et al., 1986 ; Rodda et al., 1986).

Clearly, parameters such as surface accessibility, chain termination, mobility and hydrophilicity are not independent variables but are interconnected. Attempts to single out one of these properties as a primary explanation for antigenicity and therefore to ignore other correlations may be counterproductive, since it is likely to limit our capacity to predict which regions of a protein may correspond to continuous epitopes.

The molecular dissection of protein antigens is not only of interest because it increases our understanding of immunological specificity, but also because knowledge of the antigenic structure of proteins makes it possible to manipulate the immune system and gives rise to many useful applications in molecular biology, microbiology and biotechnology. The capacity of synthetic peptides to elicit antibodies that cross-react with the corresponding complete protein is being used to develop a new generation of synthetic vaccines (Stewart & Howard, 1987) and has already produced a rich harvest of new reagents for isolating and characterizing gene products (Lerner, 1984 ; Walter, 1986).

References

Amit A.G., Mariuzza R.A., Philips S.E.V. and Poljak R.J. (1986) Three-dimensional structure of an antigen antibody complex at 2.8 Å resolution. Science 233, 747.

Atassi M.Z. (1984) Antigenic structures of proteins. Their determination has revealed important aspects of immune recognition and generated strategies for synthetic mimicking of protein binding sites. Eur. J. Biochem. 145, 1.

Barlow D.J., Edwards M.S. and Thornton J.M. (1986) Continuous and discontinuous protein antigenic determinants. Nature 322, 747.

Benjamin D.C., Berzofsky J.A., East I.J., Gurd F.R.N., Hannum C., Leach S.J., Margoliash E., Michael J.G., Miller A., Prager E.M., Reichlin M., Sercarz E.E. Smith-Gill S.J., Todd P.A. and Wilson A.C. (1984) The antigenic structure of proteins : a reappraisal. Ann. Rev. Immunol. 2, 67.

Berzofsky J.A. (1985) Intrinsic and extrinsic factors in protein antigenic structure. Science 229, 932.

Edmundson A.B. and Ely K.R. (1986) Three-dimensional analysis of the binding of synthetic chemotactic and opioid peptides in the Mcg light chain dimer. In Synthetic Peptides as Antigens, Ciba Found. Symp. 119, Wiley, Chichester, p. 107.

Fanning D.W., Smith J.A. and Rose G.D. (1986) Molecular cartography of globular proteins with application to antigenic sites. Biopolymers 25, 863.

Getzoff E.D., Geysen H.M., Rodda S.J., Alexander H., Tainer J.A. and Lerner R.A. (1987) Mechanims of antibody binding to a protein. Science (in press)

Geysen H.M., Meloen R.H. and Barteling S.J. (1984) Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. Proc. Natl. Acad. Sci. (USA) 81, 3998.

Geysen H.M., Tainer J.A., Rodda S.J., Mason T.J., Alexander H., Getzoff E.D. and Lerner R.A. (1987) Chemistry of antibody binding to a protein. Science (in press).

Hopp T.P. (1986) Protein surface analysis : Methods for identifying antigenic determinants and other interaction sites. J. Immunol. Meth. 88, 1.

Lerner R.A. (1984) Antibodies of predetermined specificity in biology and medicine. Advan. Immunol. 36, 1.

Muller S., Plaue S., Couppez M. and Van Regenmortel M.H.V. (1986) Comparison of different methods for localizing antigenic regions in histone H2A. Mol. Immunol. 23, 593.

Novotny J. and Haber E. (1986) Static accessibility model of protein antigenicity : the case of scorpion neurotoxin. Biochemistry 25, 6748.

Parker J.M.R., Guo D. and Hodges R.S. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data : correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry, 25, 5425.

Rodda S.J., Geysen H.M., Mason T.J. and Schoofs P.G. (1986) The antibody response to myoglobin - I. Systematic synthesis of myoglobin peptides reveals location and substructure of species-dependent continuous antigenic determinants. Mol. Immunol. 23, 603.

Stewart M.W. and Howard C.R. (1987) Synthetic peptides : a next generation of vaccines ? Immunol. Today 8, 51.

Tainer J.A., Getzoff E.D., Paterson Y., Olson A.J. and Lerner R.A. (1985) The atomic mobility component of protein antigenicity. Ann. Rev. Immunol. 3, 501.

Thornton J.M. and Sibanda B.L. (1983) Amino and carboxy-terminal regions in globular proteins. J. Mol. Biol. 167, 443.

Van Regenmortel M.H.V. (1984) Molecular dissection of antigens by monoclonal antibodies. In Hybridoma Technology in Agricultural and Veterinary Research (N.J. Stern & H.R. Gamble, eds), Rowman & Allanheld, Totowa, New Jersey, p. 43.

Van Regenmortel M.H.V. (1986) Definition of antigenicity in proteins and peptides. In Protides of the Biological Fluids, Ed. Peeters, H., vol. 34, Pergamon Press, Oxford, p. 81.

Van Regenmortel M.H.V., Altschuh D. and Klug A. (1986) Influence of local structure on the location of antigenic determinants in tobacco mosaic virus protein. In Synthetic peptides as antigens, Ciba Found. Symp. 119, Wiley, Chichester, p. 76.

Walter G. (1986) Production and use of antibodies against synthetic peptides. J. Immunol. Meth. 88, 149.

Westhof E., Altschuh D., Moras D., Bloomer A.C., Mondragon A., Klug A. and Van Regenmortel M.H.V. (1984) Correlation between segmental mobility and the location of antigenic determinants in proteins. Nature 311, 123.

Wilson I.A., Niman H.L., Houghten R.A., Cherenson A.R., Connolly M.L. and Lerner R.A. (1984) The structure of an antigenic determinant in a protein. Cell 37, 767.

EXHIBIT "E"

ophylline. J.

plification of neural retina S.A. 75:4451.
Lorenzetti T, atory role for etastatic human s. 25:274.
S (1979). DR s: Serological tion. J. Exp.

embrane differ- ietic and tumor lonroy A (eds): , Vol 14, ptII.

, Lloyd KO, Old ns on cultured oc. Natl. Acad.

V, Clines DB, consequences of Cancer Immunol.

Molecular Basis of Cancer, Part B: Macromolecular Recognition, Chemotherapy, and Immunology, pages 367–377

COMPUTER PREDICTION OF PROTEIN SURFACE FEATURES AND ANTIGENIC DETERMINANTS

Thomas Hopp, Ph.D.

Immunex Corporation

51 University Street, Seattle, WA 98101

For many years it has been appreciated that the arrangement of amino acids in the linear sequence of a protein is responsible for the three dimensional structure of the folded protein. However, until recently very little practical information could be obtained from amino acid sequences, because of our imperfect understanding of the way that the individual amino acids influence the conformation of a peptide chain. At present there are several useful ways of predicting conformation from sequence, including methods based on energy minimization, frequency occurrence of amino acids in particular secondary structures, and consideration of the solubility of the amino acids in aqueous and organic solvents. These methods seldom generate information that is useful on a practical level, partly because they attempt to predict too much detailed information from a given sequence. In the development of my method, I asked a simpler question, namely, is it possible to predict the locations of antibody binding sites on a protein sequence, regardless of any consideration of the precise conformation of the peptide chain? Using this criterion and a data set of twelve well known protein antigens, I developed a simple hydrophilicity analysis that reliably predicts the locations of antigenic residues in protein sequences. A listing of experiments where the outcome was predictable by my method is presented in Table I.

TABLE I

Protein Surface Features Predicted by Hydrophilicity Analysis

A. Antigenic Determinants

1. Influenza hemagglutinin: Site #1 synthetic peptide immunogen is protective in mice. Muller et al., PNAS 79, 569 (1982).
2. Influenza hemagglutinin: Sites #1, 2, and 3 contain antigenically important amino acids. Wiley et al., Nature 289, 373 (1981).
3. Influenza hemagglutinin: Sites #1 and 3 synthetic peptide immunogens are protective in mice. Shapira et al., PNAS 81, 2461 (1984).
4. Influenza hemagglutinin: Site #1 (X31 strain) is contained in a synthetic peptide immunogen recognized by T-cells. Lamb et al., Nature 300, 66 (1982).
5. Streptococcal M protein: Site #1 synthetic peptide immunogen is protective. Beachey et al., PNAS 81, 2203 (1984).
6. Poliovirus VP1: Sites #1, 3, and 5 synthetic peptide immunogens stimulate neutralizing sera. Emini et al., Nature 304, 699 (1983).
7. Poliovirus VP1: Site #3 is a neutralizing epitope. Evans et al., Nature 304, 459 (1983).
8. Poliovirus VP1: Site #3 synthetic peptide reacts with neutralizing antibodies. Wychowski et al., EMBO J. 2, 2019 (1983).
9. Foot and mouth disease virus VP1: Sites #1 and 3 of A24 strain synthetic peptide immunogens raise neutralizing antisera. Bittle et al., Nature 298, 30 (1982).
10. Foot and mouth disease virus VP1: Site #3 synthetic peptide immunogen raises neutralizing antisera. Pfaff et al., EMBO J. 1, 869 (1982).
11. Hepatitis B surface antigen: Site #1 reacts with antisera raised against HBsAg. Hopp and Woods, PNAS 78, 3824 (1981).
12. Hepatitis B surface antigen: Site #1 synthetic peptides are immunogenic in mice. Prince et al., PNAS 79, 579 (1982).
13. Hepatitis B surface antigen: Site #1 synthetic peptide is immunogenic. Bhatnagar et al., PNAS 79, 4400 (1982).
14. Hepatitis B surface antigen: Sites #2, 4, and 5 are contained in synthetic peptides that stimulate precipitating sera to HBsAg. Lerner et al., PNAS 78, 3403 (1981).
15. Hepatitis B surface antigen: Site #3 contained in synthetic peptide that stimulates anti-subtype antibodies. Dreesman et al., Nature 295, 158 (1982).
16. Hepatitis B su peptide immunog al., PNAS 80, 2
17. Influenza neura contain antigen immediately adja 303, 41 (1983).
18. Ragweed allerg important allerge Allergy and Clin
19. Herpes virus gp portion, synth antisera. Cohen
20. Histocompatibility domain is the antigenic specifi Schulze, et al.,
21. Histocompatibility domain is an all al., Biochem 22,
22. Histocompatibility synthetic immuno et al., PNAS 80,
23. Histocompatibility synthetic immuno et al., PNAS 80,
24. Beta 2 microgl monoclonal antib 6179 (1983).
25. Myelin basic pro determinant. Ha
26. Scorpion toxin Granier et al., (1984).
27. Immunoglobulin constant domain and Kehoe, Immu
28. Interf ron alpha antibody to inter 2539 (1983).
29. Interleukin 2: antibody to inter 2176 (1984).
30. Myoglobin: Sit production of ma

16. Hepatitis B surface antigen: Site #3 synthetic peptide immunogen is partially protective. Gerin et al., PNAS 80, 2365 (1983).
17. Influenza neuraminidase: Sites #1 through 4 all contain antigenically important residues or are immediately adjacent to them. Colman et al., Nature 303, 41 (1983).
18. Ragweed allergen protein RA5: Site #1 is an important allergenic determinant. Roebber et al., J. Allergy and Clin. Immunol. 71, 162 (1983).
19. Herpes virus gpD: Site #3 of the extracytoplasmic portion, synthetic peptide raises neutralizing antisera. Cohen et al., J. Virology 49, 102 (1984).
20. Histocompatibility antigen H2 K^b: Site #1 in second domain is the "gene conversion" site causing the antigenic specificity change in the H2 K^{bm1} mutant. Schulze, et al., PNAS 80, 2007 (1983).
21. Histocompatibility antigen HLA B7: Site #2 in second domain is an alloantigenic site. Lopez de Castro et al., Biochem 22, 3961 (1983).
22. Histocompatibility antigen DR alpha chain: Site #2 synthetic immunogen used to make hybridoma. Niman et al., PNAS 80, 4949 (1983).
23. Histocompatibility antigen DR beta chain: Site #2 synthetic immunogen used to make hybridoma. Niman et al., PNAS 80, 4949 (1983).
24. Beta 2 microglobulin: Site #2 recognized by a monoclonal antibody. Parham et al., J.B.C. 258, 6179 (1983).
25. Myelin basic protein: Site #1 is an encephalitogenic determinant. Hashim Immunol. Rev. 39, 60 (1978).
26. Scorpion toxin II: Sites #1 and 2 are antigenic. Granier et al., Int. J. Peptide Protein Res. 23, 187 (1984).
27. Immunoglobulin gamma chain: Site #1 of third constant domain is G1m (a) allotypic marker. Kehoe and Kehoe, Immunochemistry of Proteins 3, 87 (1979).
28. Interferon alpha 1: Site #1 synthetic peptide raises antibody to interferon. Arnheiter et al., PNAS 80, 2539 (1983).
29. Interleukin 2: Site #1 synthetic peptide raises antibody to interleukin 2: Altman et al., PNAS 81, 2176 (1984).
30. Myoglobin: Site #1 synthetic peptide causes production of macrophage inhibitory factor (MIF) by

cultured lymph n de cells. Stavitsky et al., Immunochem. 12, 959 (1975).

31. Myoglobin: Site #1 synthetic peptide used to raise a hybridoma antibody. Schmitz et al., Molec. Imm. 20, 719 (1983).
32. Cytochrome c: Site #1 causes delayed type hypersensitivity and T-cell transformation. Wang and Reichlin, Molec. Imm. 16, 805 (1979).
33. Metallothionein: Sites #1 and 3 are autoimmune antigenic sites. Winge and Garvey, PNAS 80, 2472 (1983).
34. Rous sarcoma virus transforming protein (src): Site #2 synthetic peptide immunogen raises sera that neutralize tyrosine kinase activity, cross react with yes-transforming protein (where it is Site #1) and precipitate possible cellular analogs. Gentry et al., J.B.C. 258, 11219 (1983).
35. Polyoma virus middle T antigen transforming protein: Site #1 synthetic peptide immunogen raises sera that react with middle T as well as a normal cellular protein. Ito et al., J. Virology 48, 709 (1983).

B. Interaction Sites

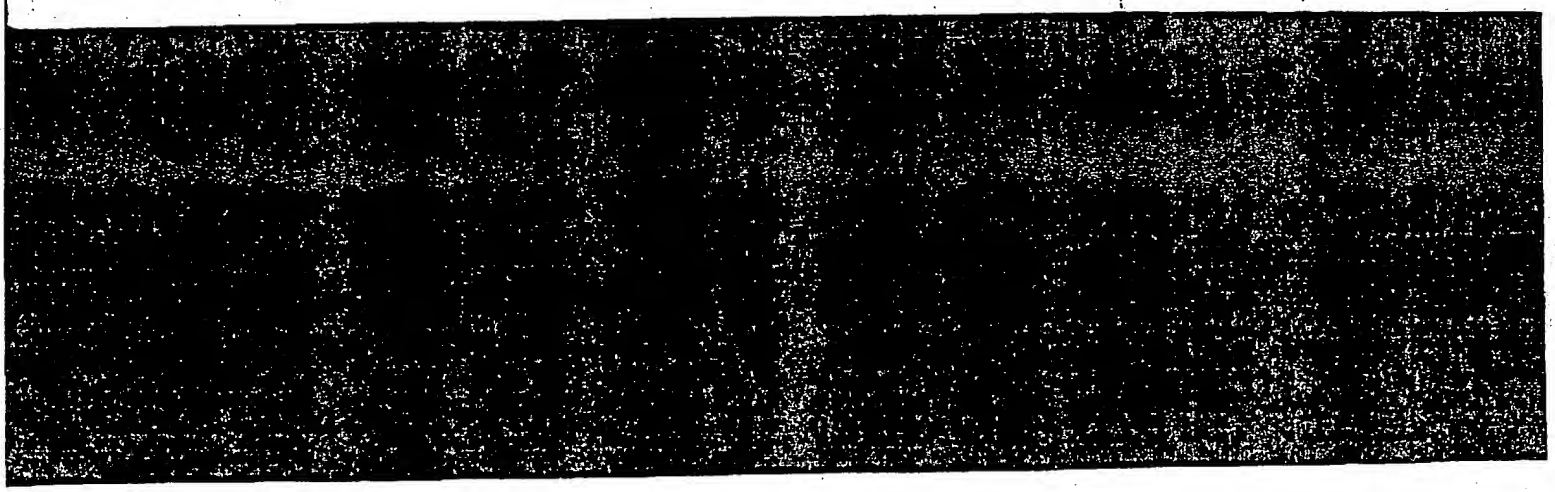
36. Immunoglobulin gamma chain: Site #1 of second constant domain is the C1q binding site. Prystowsky et al., Biochem. 20, 6349 (1981).
37. Calmodulin: Sites #3 and 4 are calcium binding sites. Waterson et al., J.B.C. 255, 962 (1980); Sasagawa et al., Biochem. 21, 2565 (1982).
38. Influenza hemagglutinin: Site #1 of X31 strain is the proteolytic processing site for cell fusion activity. Hopp and Woods, Molec. Immunol. 20, 483 (1983).
39. Fibronectin: Site #3 synthetic peptide has the cell binding activity of the whole molecule. Pierschbacher and Ruoslahti, Nature 309, 30 (1984).
40. Hepatitis B surface antigen: Site #1 contains the asparagine residue that is preferentially glycosylated over other asparagines. Peterson, J.B.C. 256, 6975 (1981).
41. Histocompatibility antigens HLA B7 and H2 K^b: Site #1 in the cytoplasmic domain contains the phosphorylatable threonine or serine residue. Pober et al., PNAS 75, 6002 (1978); Bregegere, et al., Nature 292, 78 (1981).
42. Polyoma virus middle T antigen transforming protein: Site #1 is immediately adjacent to tyrosine 315 which is phosphorylate (1984).

C. Miscellaneous

43. Fava bean lectir location of the a of the genes fo al., PNAS 76, 3:

It is clear from the variety of the ob method is highly suc problems in immunol protein chemistry. V an exhaustive list, th potential for using h including the elucida pathological organisms all of the well cha presently under inve sites have been found immunological phenom (precipitins, neutr delayed-type hype autoimmunity, enceph responses, graft rejec addition, many une surface sites have protein-protein intera binding region of imm interaction site of interaction sites of comprise locations o peptide chains, includ sites of phosphorylatic

When the wealt considered, it is clea be extremely useful ir related to carcin gene available a huge bod proteins involved in t occurring tumors and sequences, as well as of oncogenic viruses h by nucleic acid sequer is known about the from the sequences.



is phosphorylated. Hunter et al., EMBO J. 3, 73 (1984).

C. Miscellaneous

43. Fava bean lectin: Site #1 of the beta chain is the location of the annealing site for circular permutation of the genes for favin and Con A. Cunningham et al., PNAS 76, 3218 (1979).

It is clear from the number of entries in this list and the variety of the objectives achieved, that my prediction method is highly successful and has broad applicability to problems in immunology as well as the field of general protein chemistry. While no attempt was made to present an exhaustive list, the examples cited here demonstrate the potential for using hydrophilicity analysis in many areas, including the elucidation of the antigenic structures of pathological organisms. The examples in Table I include all of the well characterized major disease organisms presently under investigation. The predictable antigenic sites have been found to possess the full range of known immunological phenomena, including antibody production (precipitins, neutralizing and protective sera), delayed-type hypersensitivity, allergic responses, autoimmunity, encephalitic responses, T-cell proliferative responses, graft rejection, and lymphokine production. In addition, many unexpected examples of other protein surface sites have been listed. These include protein-protein interaction sites such as the complement binding region of immunoglobulin, the protein-cell surface interaction site of fibronectin and the protein-metal interaction sites of calmodulin. Other interaction sites comprise locations of post-translational modification of peptide chains, including proteolytic processing sites, and sites of phosphorylation and carbohydrate attachment.

When the wealth of information listed above is considered, it is clear that hydrophilicity analysis should be extremely useful in the analysis of molecular phenomena related to carcinogenesis. In particular, there has become available a huge body of sequence information on the proteins involved in transformation, both in spontaneously occurring tumors and in virally induced tumors. These sequences, as well as the genomic sequences of a variety of oncogenic viruses have been obtained for the most part by nucleic acid sequencing, and therefore little or nothing is known about the structures of the proteins produced from the sequenc s. In this regard it is instructive to

consider several examples of experiments suggested by hydrophilicity analysis of tumor virus genome encoded proteins.

The first example is an investigation of the hydrophilicity properties of the two protein products of the env region of the recently described adult thymic leukemia virus (ATLV) genome. In figure 1, the hydrophilicity profile for the heavy chain of the env translation product is compared to two other viral envelope proteins, those of the influenza virus (HA1) and the hepatitis B virus (HBsAg). This comparison seems especially interesting because it has been noted that the heavy and light chains of the retroviral env translation products bear a general resemblance to the two chains of the influenza hemagglutinin (HA1 and HA2). A comparison of the ATLV env light chain to HA2 and to a membrane glycoprotein product of the early region of adenovirus (ADVE16) is shown in figure 2. The most striking similarity among these plots is between the HA1 and env1 heavy chains. A number of common features lead me to conclude that these two proteins share closely similar three dimensional structures, even though they have not been reported to be homologous.

The HA1 and env1 chains are obviously similar in length, although env1 is slightly shorter. More significantly, the hydrophilicity profiles show a great number of common characteristics. Each has a broad hydrophobic valley near the N-terminus. This region of HA1 makes up the central strand of the membrane associated globular domain. At their C-termini, the two proteins again show a similar feature in the large terminal peak. This feature is associated with the known proteolytic processing site of the influenza hemagglutinin and the proposed processing site of the env product. The profile for HBsAg is included in figure 1 to show that not all envelope proteins share such hydrophilicity profiles. While HBsAg does have an N-terminal hydrophobic valley, it clearly lacks a C-terminal peak. It is also obviously different in being much shorter than HA1 or env1, and in having a broad central hydrophobic valley that may actually be a membrane spanning segment. Although HA1 and env1 both have a number of hydrophobic valleys in their central regions, neither has one of sufficient length to span a membrane. This central region of the HA1 chain is known to comprise the globular domain at the distal end of the hemagglutinin spike and to contain the binding site

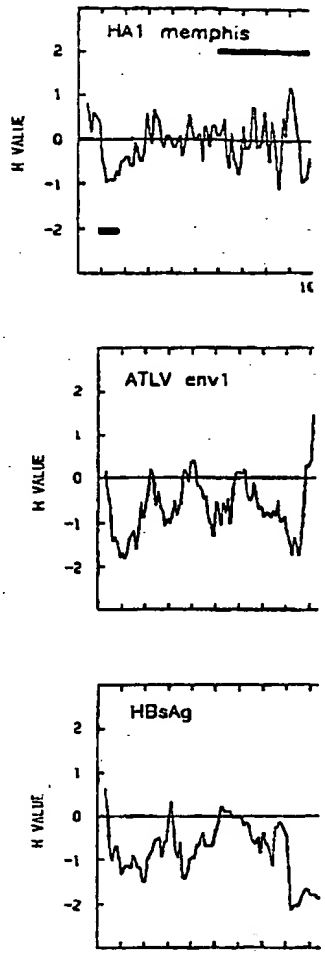
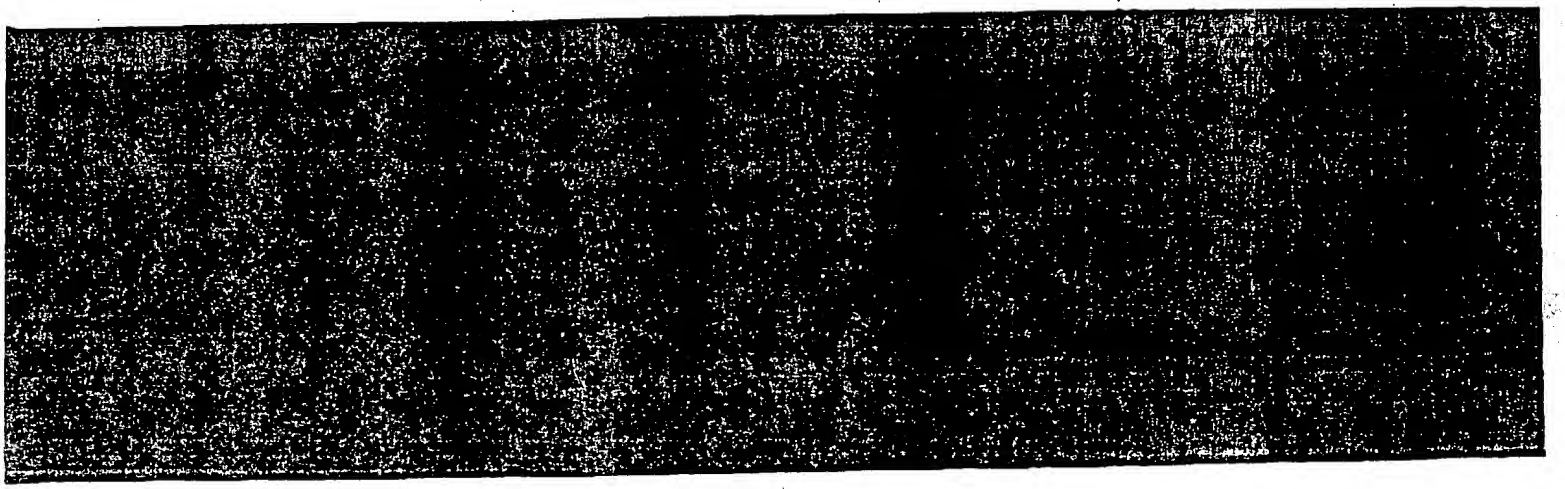


Figure 1. Hydropl antigens. The bar ab extent of the globul profile, the solid bars bars represent helices



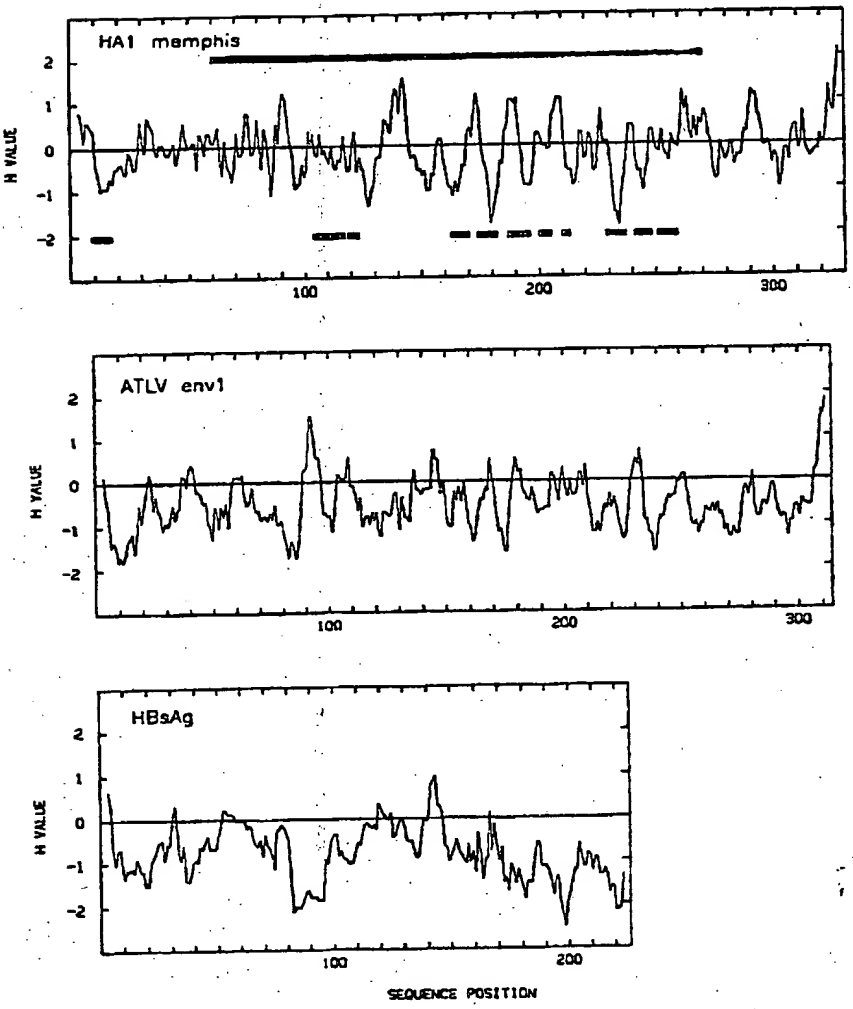


Figure 1. Hydrophilicity analysis of viral surface antigens. The bar above the profile for HA1 indicates the extent of the globular cell-binding domain. Below the profile, the solid bars represent β-strands and the hatched bars represent helic s.

suggested by
nome encoded

ıtion of the
n products of
adult thymic
gure 1, the
ı of the env
viral envelope
HA1) and the
parison seems
ıoted that the
ınv translation
two chains of
A comparison
to a membrane
of adenovirus
most striking
HA1 and env1
·es lead me to
y similar three
have not been

ısly similar in
ıorter. More
show a great
has a broad
This region of
the membrane
rmini, the two
large terminal
h the known
ı hemagglutinin
product. The
ı show that not
ilicity profiles.
ophobic valley,
also obviously
ır env1, and in
lley that may
Alth ugh HA1
obic vall ys in
ufficient length
the HA1 chain
t the distal end
he binding site

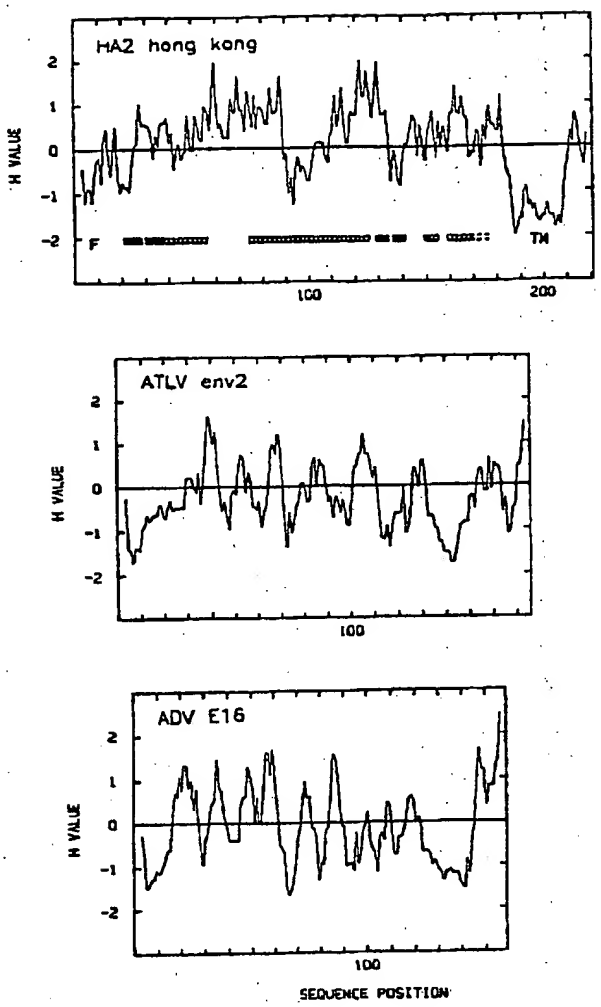
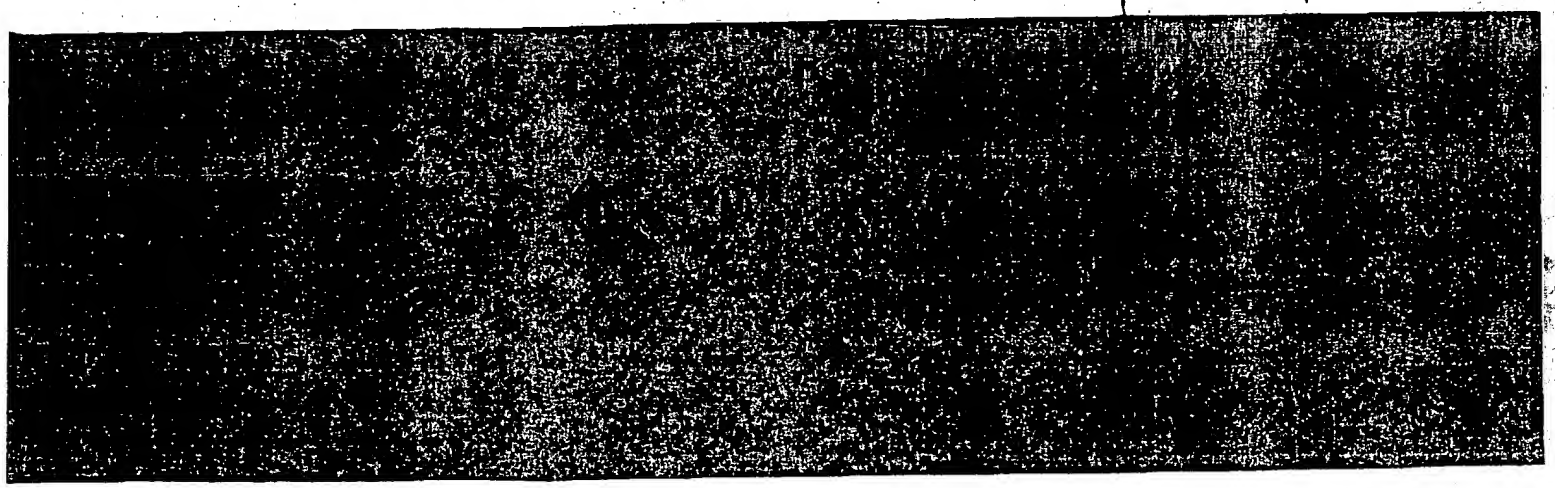
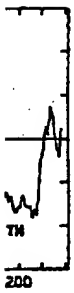


Figure 2. Hydrophilicity analysis of surface antigen light chains. In the HA2 plot, solid bars represent β-strands and hatched bars represent helices; F, membrane fusion region; TM, transmembrane anchoring segment.

for cell-surface sialic
domain is formed into
with several associat
repeating secondary
hydrophilicity profile
valleys in the central
are related to the d
represent the highly
the strands. In this
that the env1 hydrop
large peaks and valle
that it, too contair
repeating β-strands.
peaks and valleys of
significant differences
It is not possible, fr
favor of any homology
these two proteins, h
similarities, it would
evolved from a comm
these two proteins
dimensional structu
determinants for the
located on the hydr
central portion of the
acids 90-95, 142-147,
third, and fourth l
explored for usefulr
against this disease.
homologous portions
AIDS associated viru
vaccines against that

In figure 2, sev
env2 and HA2.
hydrophobic valley, w
fusion activity in
hydrophobic stretches
membrane anchoring
may be substantial d
HA2 is significantly l
the short-period spik
content of the molec
have less of this hel
the short-period spi
peaks and valleys th
β-strands. Interesti





for cell-surface sialic acid. The core of this cell binding domain is formed into an eight stranded β pleated sheet with several associated short helical stretches. These repeating secondary structures are reflected in the hydrophilicity profile by the series of large peaks and valleys in the central part of the plot. Most of the valleys are related to the different β strands, while the peaks represent the highly exposed chain turns at the ends of the strands. In this light, it is most interesting to note that the env1 hydrophilicity plot also shows a series of large peaks and valleys in its central region, suggesting that it, too contains a globular domain composed of repeating β-strands. It is not possible to align all of the peaks and valleys of HA1 and env1, so there must be significant differences between the two proteins as well. It is not possible, from these findings, to make a case in favor of any homology or evolutionary relationship between these two proteins, however, given the number of general similarities, it would not be surprising if they had indeed evolved from a common ancestor. Most significantly, if these two proteins do share similar overall three dimensional structures, then important antigenic determinants for the neutralization of ATL virus must be located on the hydrophilic peaks found throughout the central portion of the env1 protein. In particular, amino acids 90-95, 142-147, and 230-235, comprising the second, third, and fourth highest peaks for env1 should be explored for usefulness as synthetic peptide vaccines against this disease. Past experience indicates that the homologous portions of the env products of the related AIDS associated viruses should also be considered as vaccines against that disease.

In figure 2, several similarities are apparent between env2 and HA2. Each protein has an N-terminal hydrophobic valley, which is associated with the membrane fusion activity in HA2. Both proteins have long, hydrophobic stretches near their C-termini, likely to be membrane anchoring regions. It is also clear that there may be substantial differences between the two, because HA2 is significantly longer, and its profile shows more of the short-period spikiness associated with the large helix content of the molecule. The shorter env2 protein may have less of this helix, although it does contain some of the short-period spikes in addition to the pronounced peaks and valleys that may imply a greater content of β-strands. Interestingly, a more convincing structural

antigen light
sent β-strands
mbrane fusion
it.

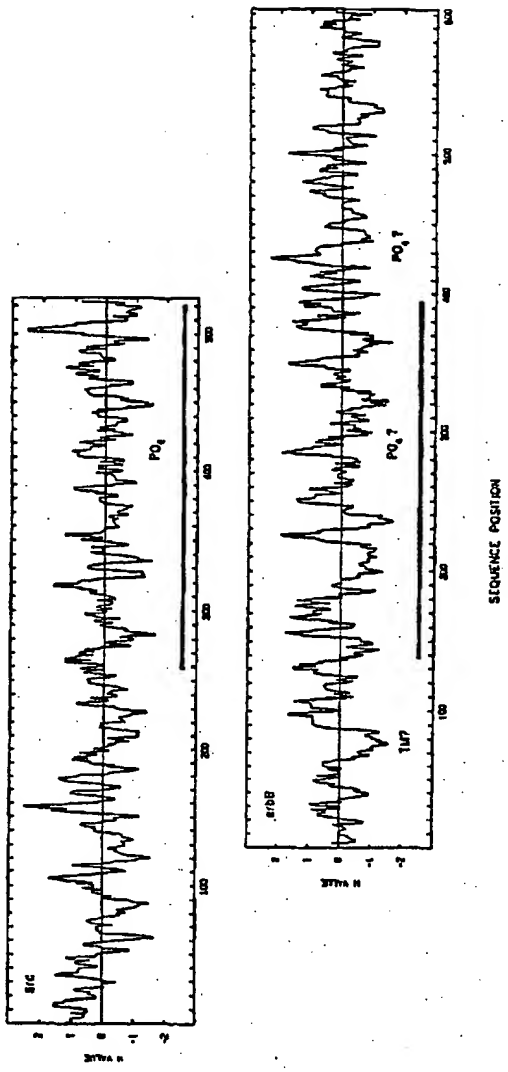


Figure 3. Hydrophilicity analysis of src and erbB oncogene proteins. The horizontal bars represent regions of shared homology between src, erbB, protein kinase, and EGF receptor. PO_4 indicates a site of known or proposed tyrosine phosphorylation; TM, proposed transmembrane segment.

similarity is suggestec
that for adenovirus
protein has not bee
protein, it is intriguir
the hydrophilicity plo
of two very different

Examples of h
important family of ca
figure 3. The two re
erbB each contain a r
protein kinase, and to
growth factor recep
synthetic peptide comp
peak for src (residue
neutralize its tyrosine
the homolgous yes t
normal cellular analog
raise antibodies spec
(residues 155 to 160) l
not homologous to th
Such an antiserum sh
function of this regior
should cross react str
cellular c-src gene p
region, but may be no
many amino acid subs
that studies of the a
erbB protein would als
site #1 has been pr
phosphorylation in erb
tyrosine phosphorylati
two sites would clarify
in erbB, and, because
highly conserved bet
same antibodies shoul
well. Another regio
N-terminus, where t
proposed to reside
membrane. Antibodie
raised using synthetic
the hydrophilic reg
antibodies and the o
possible to begin a m
group of transforming

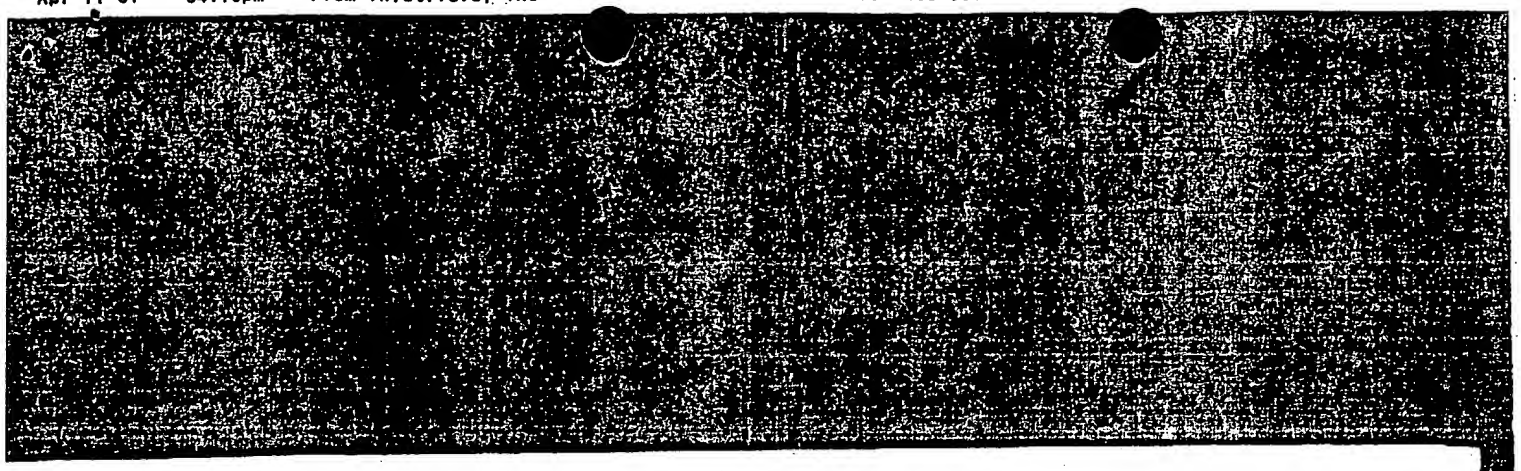


Figure 3. Hydrophilicity analysis of src and erbB oncogene proteins. The horizontal bars represent regions of shared homology between src, erbB, protein kinase, and EGF receptor. PO_4 indicates a site of known or proposed tyrosine phosphorylation; TM, proposed transmembrane segment.

similarity is suggested by comparing the plot for env2 to that for adenovirus E16. Although the E16 membrane protein has not been implicated as a viral structural protein, it is intriguing to note the apparent similarity of the hydrophilicity plots for membrane associated proteins of two very different types of oncogenic viruses.

Examples of hydrophilicity analysis of another important family of cancer related proteins are shown in figure 3. The two retroviral oncogene products, src and erbB each contain a region of homology to each other, to protein kinase, and to the cytoplasmic portion of epidermal growth factor receptor. Antibodies raised against a synthetic peptide comprising the second highest prediction peak for src (residues 498 to 512) have been shown to neutralize its tyrosine kinase activity, and to precipitate the homolgous yes transforming protein as well as the normal cellular analog of src. It would be interesting to raise antibodies specific to the highest peak for src (residues 155 to 160) because this region of the molecule is not homologous to the erbB product or EGF receptor. Such an antiserum should be useful in characterizing the function of this region of the src molecule and because it should cross react strongly with the recently characterized cellular c-src gene product, which is identical in this region, but may be non-cross reactive with yes, which has many amino acid substitutions at this site. It is likely that studies of the antigenic peptides predicted for the erbB protein would also be quite useful, because predicted site #1 has been proposed as a major site of tyrosine phosphorylation in erbB, and site #3 is homologous to the tyrosine phosphorylation site in src. Antibodies to these two sites would clarify the role of tyrosine phosphorylation in erbB, and, because the amino acids around site #1 are highly conserved between erbB and EGF receptor, the same antibodies should cross react with the receptor as well. Another region of itnerest on erbB is at the N-terminus, where the first 60-70 residues have been proposed to reside on the outside of the cell plasma membrane. Antibodies specific to this region could be raised using synthetic peptides comprising one or more of the hydrophilic regions found there. Using these antibodies and the ones described above it should be possible to begin a molecular dissection of this important group of transforming proteins.

EXHIBIT "F"

Brief Definitive Report

T CELL CLONES SPECIFIC FOR AN AMPHIPATHIC α-HELICAL REGION OF SPERM WHALE MYOGLOBIN SHOW DIFFERING FINE SPECIFICITIES FOR SYNTHETIC PEPTIDES

A Multiview/Single Structure Interpretation of Immunodominance

By KEMP B. CEASE, IRA BERKOWER, JENA YORK-JOLLEY, AND JAY A. BERZOFSKY

From the Metabolism Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892; and the Bureau of Biologics, Food and Drug Administration, Bethesda, Maryland 20892

Characterization of immunodominant T cell sites has been effectively performed by a number of laboratories (reviewed in 1) using protein sequence variants, cleavage fragments, and synthetic peptides. Some studies (2) have been interpreted as supportive of the possibility of multiple conformations of peptide antigen and/or multiple antigen-binding sites on the Ia molecule. The possibility that distinct T cell specificities might reflect distinct recognition or views of a single peptide conformation associated with a single Ia site has received little attention, primarily because little could be inferred about the conformation of the antigenic peptide on the APC in most experimental systems studied.

We have previously described an immunodominant site in sperm whale myoglobin in a region encompassing glutamic acid 109, identified using myoglobin sequence variants (3), and have subsequently isolated T cell clones with the same reactivity pattern (4). In this paper we characterize two such clones using a panel of synthetic peptides. The clones showed different response patterns that are found to be totally consistent with a model of the distinct T cell specificities reflecting distinct "views" of an amphipathic α-helical conformation. Thus, when one considers the likely secondary structure of antigen existing in association with the complex structure of the Ia molecule, distinct T cell recognition specificities need not imply distinct structural forms of antigen or sites of antigen binding, but rather may reflect distinct views recognized by the T cell receptor.

Materials and Methods

Mice. B10.D2 and (B10.D2 × B10.BR)F_1 mice were obtained from The Jackson Laboratory (Bar Harbor, ME).

T Cell Clones. T cell clone 9.27 was derived from B10.D2 mice as described (4). T cell clone 1.2 was derived independently from (B10.D2 × B10.BR)F_1 mice as described and has been referred to as F_1(D2)1.2 in previous studies (5). Both clones are specific for the glutamic acid 109 region of sperm whale myoglobin and are restricted to I-A^d (5).

Synthetic Peptides. Synthetic peptides of sperm whale myoglobin 102–118, 104–118, 106–118, 108–118, 109–118, 110–118 were synthesized by manual solid-phase peptide

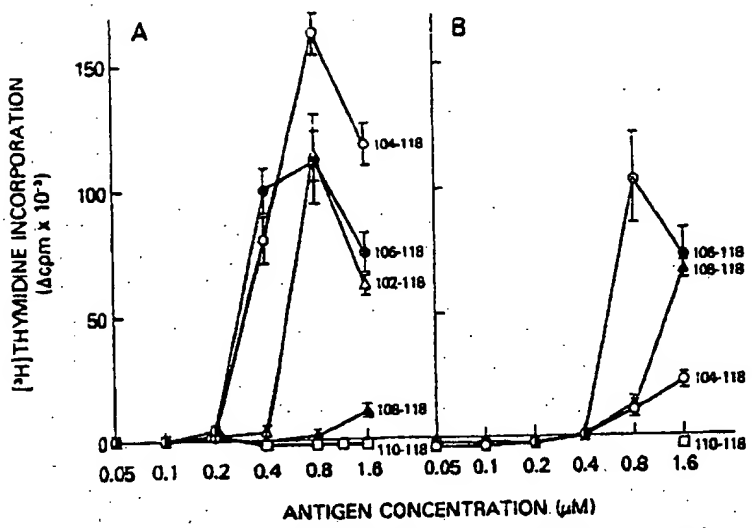


FIGURE 1. Proliferative responses of T cell clones 1.2 and 9.27 to synthetic myoglobin peptides corresponding to residues 102–118 (Lys-Tyr-Leu-Glu-Phe-Ile-Ser-Glu-Ala-Ile-Ile-His-Val-Leu-His-Ser-Arg) and portions thereof presented by H-2^d APCs. Assays were performed as described in the text. Background-subtracted geometric means are shown with SEM confidence intervals. (*A*) clone 1.2, (*B*) clone 9.27. The background thymidine incorporation without antigen was 2,298 cpm for clone 1.2 and 3,493 cpm for clone 9.27.

synthesis using a modification of the method of Corley et al. (6) and purified to homogeneity by gel filtration on BioGel P4 followed by reversed-phase HPLC. Concentration determination and composition confirmation were determined by amino acid analysis kindly performed by Robert Boykins (Food and Drug Administration).

T Cell Proliferation Assay. Assays were performed as described previously (4, 5).

Results and Discussion

Previous studies (4, 5) using sequence variants of native myoglobin had shown that the sites in native sperm whale myoglobin seen by these clones included glutamic acid 109 and possibly histidine 116. Available sequence variants do not enable higher resolution analysis by this approach. Thus, we synthesized the peptides in the nested series from 102–118 to 110–118. Fig. 1, *A* and *B* show peptide dose-response curves for clones 1.2 and 9.27. Both clones respond well to 102–118 and 106–118 (Fig. 1 and Table I). However, clone 1.2 responds much better to 104–118 than to 108–118, whereas the reverse is true for clone 9.27. The rank order of potency for clone 9.27 is not a simple function of peptide length, as 104–118 is less potent than the shorter peptides 106–118 and 108–118 even though it contains all of the sequence present in these latter peptides. The decrease in activity from peptide 106–118 to peptide 104–118 is then reversed when the peptide is further lengthened to 102–118. This result further indicates that activity is not simply a function of peptide length. Though not shown in this experiment, the potency of peptide 102–118 for stimulating clone 9.27 was consistently greater than or equal to that of peptide 106–118. Neither clone responded to 110–118. In subsequent experiments 109–118 was found to be inactive for both clones (Table I).

TABLE I

Proliferative Response of Clones 9.27 and 1.2 to Peptides 109–118 and 102–118

Clone	Peptide	Geometric mean [^{3}H]thymidine incorporation (cpm) at an antigen concentration of:		
		0 μM	1.0 μM	3.2 μM
9.27	109–118		296 (1.15)*	257 (1.97)
	102–118		151,060 (1.03)	119,378 (1.09)
	No antigen	260 (1.27)		
1.2	109–118		570 (1.66)	238 (1.35)
	102–118		169,457 (1.03)	173,410 (1.05)
	No antigen	402 (1.52)		

Assays were performed as described in the text.
* Geometric means are shown with the SEM factor in parentheses.

Thus, the identification of the region around residue 109 as the immunodominant site was confirmed using synthetic peptides, and residues on both sides of 109 in the sequence appear to contribute to antigenicity. While these data show a consensus segment from 106–118 for stimulation of the clones, they also reveal a differential response pattern to longer and shorter peptides.

This segment folds into a highly amphipathic α helix in native myoglobin with the hydrophobic residues on one face and the hydrophilic on the opposite face (Fig. 2*A*). Refolding of this peptide into this conformation in the hydrophobic/hydrophilic interface at the surface of the presenting cell should be energetically favored. The data we present here on the 102–118 region, along with our previously reported data (7a) on the 132–146 site of sperm whale myoglobin, led us to hypothesize (1) that the amphipathic helix may be a general feature common to many immunodominant T cell sites. Indeed, a sequence consistent with formation of an amphipathic helix is seen in the majority of T cell antigenic sites described to date. Thus, we suggest that such structures may frequently represent an integral part of the stimulation complex for the T cell receptor.

If in fact an α-helical antigen conformation is presented to and recognized by the T cell receptor, simple end effects and folding patterns that differentially affect one or the other T cell clone recognition regions would appear likely (Fig. 2*B*). For instance, clone 1.2, which is more sensitive to end effects at 108, that is to say, does not respond to peptide 108–118, may include this residue in its epitope (see Fig. 2*B*). In contrast, the *N*-terminal limit of the epitope recognized by clone 9.27 may be Glu 109 itself, so that adding on just one more residue at position 108, to mask the α-amino group of 109, is sufficient to stimulate the clone. Thus, these data are consistent with a model of multiple T cell specificities arising from multiple views of a single antigen conformation at a single Ia-binding site and do not require postulation of multiple conformations or binding sites.

Multiple distinct functional sites on each Ia molecule have been proposed (2, 8–13) to explain the results of antibody blocking and Ia mutant studies, as well as to account for findings of distinguishable T cell specificities for a given antigenic site. Among these, elegant studies by Allen *et al.* (2, 13) have focused on the hen egg lysozyme (HEL) system using molecular variations in antigen and Ia to probe the specificity of a panel of T cell hybridomas. Two T cell hybridomas specific for HEL 46-61 in association with I-A^k, but differing in fine specificity,

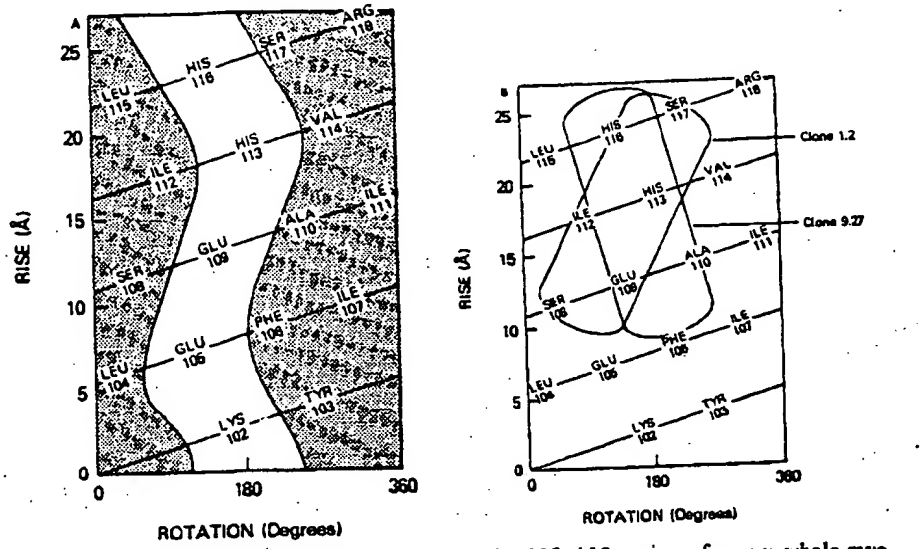


FIGURE 2. Secondary structure presentation for the 102–108 region of sperm whale myoglobin. In these α-helical net displays the cylinder of the helix is "cut" longitudinally, opened, and folded flat (7). (*A*) The amphipathic character of this segment. Hydrophobic residues are shaded and tend to fall on one face of the helix while the hydrophilic residues are positioned on the opposite face. (*B*) Hypothetical T cell receptor recognition envelopes for clones 1.2 and 9.27 (see text).

were found (13, 14) to be differentially sensitive to two mutations in A^k. As in the case of the bm12 mutant (9, 10), these results could be interpreted equally well in terms of two disjoint sites on Ia or a single Ia site recognized differently by two T cells that are differentially sensitive to these substitutions. The recent finding that these mutations are only three residues apart (14; Allen, P., and D. McKean, personal communication) supports the latter interpretation. Given that Ia, a member of the Ig gene superfamily, shares homology and domain structure with Fab, it is not unreasonable to suppose that the several hypervariable regions may cluster to form a single combining site for antigen. Indeed, the fact that a single mutation at position 67 resulted in loss of stimulation of T cell hybridomas specific for at least three distinct peptides of HEL representing over half of the cases characterized (13) further supports a single site. In the other cases, antigen could still be in the same conformation at the same site on Ia but viewed by the T cell slightly differently in association with different adjacent Ia residues. The intriguing difference in rank order of potency of different antigenic peptides for two hybridomas (2) can be explained by the presence of a turn at residues 55–56, as in native HEL, such that additional residues introduce stimulatory information for one clone but only steric hinderance for the other. Other studies (15–17) have emphasized the importance of distinct T cell specificities for a common antigenic site but have not suggested structural interpretations. These cases are, however, also consistent with a multiview model of T cell recognition of antigen in a single conformation at a single site. Though all of the data, including our own, are also consistent with models proposing multiple antigen conformations and multiple Ia antigen–binding sites, we would argue that the

existing data do not necessitate postulating these more complex models. Thus, distinct T cell recognition specificities need not imply distinct structural forms of antigen or sites of antigen binding but rather may reflect distinct views recognized by the T cell receptor.

The finding of differing T cell clone fine specificities within an immunodominant site additionally suggests that immunodominance represents the focusing of a polyclonal response on a limited region of the antigen and is not a simple monoclonal T cell expansion. Other findings are also consistent with this interpretation (2, 15–17; Livingstone, A., J. Rothbard, and C. G. Fathman, personal communication).

Summary

The T cell response to sperm whale myoglobin in the *H-2*^{d} haplotype has been shown to be largely focused on a limited region around glutamic acid 109 recognized in association with I-A^d. T cell clones 9.27 and 1.2 have been previously (4, 5) shown to reflect this specificity and MHC restriction. In this study we have used a panel of synthetic peptides from the region 102–118 of myoglobin to characterize the specificities of these representative clones. The segment from 106–118 was found to represent a consensus region for recognition by both clones. However, we saw significant differences between clones in the hierarchy of responsiveness to peptides within the panel. In as much as the peptide and the I-A^d molecule remain constant, these differences derive from differences in how each T cell receptor interacts with the antigen. This peptide segment is an amphipathic α helix in native myoglobin, meaning that one side is hydrophobic and the other hydrophilic. It is one of the prototype cases that led us to find that amphipathic helices constitute the majority of immunodominant sites recognized by helper T cells (1). It is likely that the peptide will refold into an amphipathic helix stabilized by the interface at the surface of the presenting cell. When such secondary conformation is considered, these data are consistent with a model of multiple T cell specificities arising from multiple views of a single antigen conformation at a single Ia-binding site and do not require postulation of multiple conformations or binding sites.

Additionally, the finding of distinct specificities suggests that the immunodominance of this site depends not on the dominance of a single clone, but on the focusing of a polyclonal response on a single region of the molecule in association with I-A^d. The immunodominance of this particular region of the protein may thus depend on intrinsic features of the site, such as potential to form an amphipathic helix, as well as extrinsic factors such as binding properties of the I-A molecule.

We are grateful to Dr. Richard Hodes for critical reading of the manuscript and to Dr. Hodes and Dr. Alfred Singer for helpful discussion.

Received for publication 2 June 1986 and in revised form 31 July 1986.

References

1. DeLisi, C., and J. A. Berzofsky. 1985. T-cell antigenic sites tend to be amphipathic structures. *Proc. Natl. Acad. Sci. USA.* 82:7048.

2. Allen, P. M., G. R. Matsueda, E. Haber, and E. R. Unanue. 1985. Specificity of the T c ll receptor: two different determinants are generated by the same peptide and the I-A^k molecule. *J. Immunol.* 135:368.
3. Berkower, I., G. K. Buckenmeyer, F. R. N. Gurd, and J. A. Berzofsky. 1982. A possible immunodominant epitope recognized by murine T lymphocytes immune to different myoglobins. *Proc. Natl. Acad. Sci. USA.* 79:4723.
4. Berkower, I., L. A. Matis, G. K. Buckenmeyer, F. R. N. Gurd, D. L. Longo, and J. A. Berzofsky. 1984. Identification of distinct predominant epitopes recognized by myoglobin-specific T cells under the control of different *Ir* genes and characterization of representative T cell clones. *J. Immunol.* 132:1370.
5. Berkower, I., H. Kawamura, L. A. Matis, and J. A. Berzofsky. 1985. T cell clones to two major T cell epitopes of myoglobin: effect of I-A/I-E restriction of epitope dominance. *J. Immunol.* 135:2628.
6. Corley, L., D. H. Sachs, and C. B. Anfinsen. 1972. Rapid solid-phase synthesis of bradykinin. *Biochem. Biophys. Res. Commun.* 47:1353.
7. Dunnill, P. 1968. The use of helical net-diagrams to represent protein structures. *Biophys. J.* 8:865.

7a. Berkower, I., G. K. Buckenmeyer, and J. A. Berzofsky. 1986. Molecular mapping of a histocompatibility-restricted immunodominant T cell epitope with synthetic and natural peptides: implications for T cell antigenic structure. *J. Immunol.* 136:2498.

8. Burger, R., and E. M. Shevach. 1980. Monoclonal antibodies to guinea pig Ia antigens. II. Effect on alloantigen-, antigen-, and mitogen-induced T lymphocyte proliferation in vitro. *J. Exp. Med.* 152:1011.
9. Lin, C.-C. S., A. S. Rosenthal, H. C. Passmore, and T. H. Hansen. 1981. Selective loss of antigen-specific *Ir* gene function in *IA* mutant B6.C-*H-2*^{bm12} is an antigen presenting cell defect. *Proc. Natl. Acad. Sci. USA.* 78:6406.
10. Beck, B. N., P. A. Nelson, and C. G. Fathman. 1983. The I-A^b mutant B6.C-H-2^{bm12} allows definition of multiple T cell epitopes on I-A molecules. *J. Exp. Med.* 157:1396.
11. Needleman, B. W., M. Pierres, C. A. Devaux, P. N. Dwyer, A. Finnegan, D. H. Sachs, and R. J. Hodes. 1984. An analysis of functional T cell recognition sites on I-E molecules. *J. Immunol.* 133:589.
12. Cohn, L. E., L. H. Glimcher, R. A. Waldmann, J. A. Smith, A. Ben-Nun, J. G. Seidman, and E. Choi. 1986. Identification of functional regions on the I-A^b molecule by site-directed mutagenesis. *Proc. Natl. Acad. Sci. USA.* 83:747.
13. Allen, P. M., D. J. McKean, B. N. Beck, J. Sheffield, and L. H. Glimcher. 1985. Direct evidence that a class II molecule and a simple globular protein generate multiple determinants. *J. Exp. Med.* 162:1264.
14. Brown, M. A., L. A. Glimcher, E. A. Nielsen, W. E. Paul, and R. N. Germain. 1986. T-cell recognition of Ia molecules selectively altered by a single amino acid substitution. *Science (Wash. DC).* 231:255.
15. Manca, F., J. A. Clarke, A. Miller, E. E. Sercarz, and N. Shastri. 1984. A limited region within hen egg-white lysozyme serves as the focus for a diversity of T cell clones. *J. Immunol.* 133:2075.
16. Shastri, N., A. Oki, A. Miller, and E. E. Sercarz. 1985. Distinct recognition phenotypes exist for T cell clones specific for small peptide regions of proteins. Implications for the mechanisms underlying major histocompatibility complex–restricted antigen recognition and clonal deletion models of immune response gene defects. *J. Exp. Med.* 162:332.
17. Shimonkevitz, R., S. Colon, J. W. Kappler, P. Marrack, and H. M. Grey. 1984. Antigen recognition by H-2-restricted T cells. II. A tryptic ovalbumin peptide that substitutes for processed antigen. *J. Immunol.* 133:2067.

0022-1767/87/1381-0204$02.00/0
THE JOURNAL OF IMMUNOLOGY
Vol. 138, 204-212, No. 1, January 1, 1987
Printed in U.S.A.

STRONG CONFORMATIONAL PROPENSITIES ENHANCE T CELL ANTIGENICITY

JOHN L. SPOUGE,* H. ROBERT GUY,* JAMES L. CORNETTE* HANAH MARGALIT,* KEMP CEASE,† JAY A. BERZOFSKY,† AND CHARLES DELISI*

*From the *Laboratory of Mathematical Biology and †Metabolism Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892*

The ability to predict T cell antigenic peptides would have important implications for the development of artificial vaccines. As a first step towards prediction, this report uses a new statistical technique to discover and evaluate peptide properties correlating with T cell antigenicity. This technique employs Monte Carlo computer experiments and is applicable to many problems involving protein or DNA.

The technique is used to evaluate the contribution of various peptide properties to helper T cell antigenicity. The properties investigated include amphipathicities (α and β), conformational propensities (α, β, turn and coil), and the correlates of α-helices, such as the absence of helix-breakers and the positioning of the residues which stablize α-helical dipoles. We also investigate segmental amphipathicity. (A peptide has this property when it contains at least two disjoint subpeptides, one hydrophobic, one hydrophilic.) Statistical correlations and stratifications assessed independent contributions to T cell antigenicity.

The findings presented here have important implications for the manufacture of peptide vaccines. These implications are as follows: if possible, peptide vaccines should probably be those protein segments *a*) which have a propensity to form amphipathic α-helices, *b*) which do not have regions with a propensity to coil conformations, and *c*) which have a lysine at their COOH-terminus. The last two observations are of particular use in manufacturing peptides vaccines: they indicate where the synthetic peptides should be terminated. These implications are supported by the findings given below.

The significances (p values) support the following statistical generalites about antigenic conformations: *1*) most helper T cell antigenic sites are amphipathic α-helices; *2*) α-helical amphipathicity and propensity to an α-helical conformation contribute independently to T cell antigenicity; *3*) there is evidence that some T cell antigenic sites are β conformations instead of α-helices; *4*) T cell antigenic sites avoid random coiled conformations; and *5*) T cell antigenic sites are usually *not* segmentally amphipathic.

α-Helical amphipathicity was significant, but segmental amphipathicity was not. This has implications for the dimensions of the structure interacting with the hydrophobic portion of an amphipathic T cell antigenic site.

Lysines are unusually frequent at the COOH-terminal of T cell antigenic sites, even after accounting for tryptic digests. These lysines can stabilize α-helical peptides by a favorable interaction with α-helical dipoles. This interaction, which occurs with other charged residues and not just lysine, is probably stronger in peptides than in native proteins because of the terminal backbone charges in free peptides. This stabilization may explain why alteration of COOH-terminal lysines often destroys antigenic activity: this experimental fact, never before noted as a general observation, is predicted by our theory.

Our statistics are consistent with a "conformational hypothesis": helper T cell immunodominant sites tend to be peptides with strong conformational propensities that stabilize under hydrophobic interaction with some structure on the antigen-presenting cell, possibly a class II major histocompatibility complex protein. The conformational hypothesis is an extension of the amphipathicity hypothesis, which does not consider conformational propensities. Because small peptides do not commonly take stable conformations, our results support the quite reasonable notion that immunodominant sites are often those peptides most able to present the T cells with a consistent conformational picture.

The studies presented here detect several properties of the amino acid sequences of antigenic peptides which correlate with helper T cell immunodominance. These properties suggest fundamental chemical rules governing T cell recognition of antigens. In addition, these properties would be valuable for incorporation into the rational design of any synthetic vaccine.

Prediction of T cell antigenic peptides would have important implications for the development of artificial vaccines. Such vaccines would be particularly useful in diseases such as leprosy, caused by organisms which are hard to culture and for which the cellular arm of the immune system is the principal defense. Even when antibody production is the primary goal of vaccination, a secondary or anamnestic response requires the induction of helper T cell immunity. Prediction of peptides for use as vaccines requires discovery and confirmation of properties correlating with T cell antigenicity. The purpose of this report is to find such properties for the case of helper T cells.

Received for publication June 26, 1986.
Accepted for publication September 13, 1986.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

There are, for our purposes, two antigenic challenges which raise a T cell response: a) challenge by native protein (N), and b) challenge by a short peptide segment (P) produced either by synthesis or by cleavage from the native protein. There are four possibilities for primary, followed by secondary, challenge: NN, NP, PN, and PP. Although artificial vaccines are PN (peptide vaccination should immunize against native protein), most laboratory experiments are either NN or NP.

NN and NP experiments can localize immunodominant sites, i.e., minimal peptides within a native protein which are the focus of T cell response. This report will call the experimental peptides containing the immunodominant sites *antigenic sites*. ("antigenicity" in this paper always refers to T cell antigenicity.) If the NN or NP immunodominant sites were good PN sites, they would be candidate peptides for artificial vaccines. Apart from pure scientific interest, this hypothesis motivates determination of immunodominant sites. Immunodominant sites usually are found by trials using many peptides; systematic prediction could speed their investigation considerably.

In vivo, an antigenic protein probably passes through the following three main steps before raising a helper T cell response. *1*) Processing: an antigen-presenting cell (APC),[1] usually a macrophage, dendritic cell, or B cell, ingests the protein and then digests it into smaller peptides (1, 2). *2*) Presentation: these peptides are then presented to T cells, probably in conjunction with a class II major histocompatibility complex (MHC) protein on the APC surface (3–5). *3*) Recognition: a helper T cell receptor then recognizes some combination of peptide and class II protein, and initiates a T cell response.

Two antigenic properties are currently thought to contribute to this process: amphipathicity and α-helicity.

A structure is amphipathic when it has both a hydrophobic portion and a hydrophilic portion (6). A peptide is segmentally amphipathic when the peptide contains at least two disjoint subpeptides, one hydrophobic, the other hydrophilic. We call a peptide α-amphipathic if, when the peptide is put into an α-helical conformation, one side of the α-helix is hydrophobic, the other side hydrophilic. Both segmental amphipathicity and α-amphipathicity are believed to contribute to T cell antigenicity, although opinions about their relative importance differ (7–10).

Present evidence suggests that some antigens assume α-helical conformations (11, 12) which are then stabilized by hydrophobic interactions with a class II protein on the APC (13–15). According to the amphipathicity hypothesis (7), these antigens would tend to be α-amphipathic, with the hydrophobic side of the helix interacting with the APC, and the hydrophilic side with the T cell.

Much is known about the α-helical conformation. Certain amino acids are helix-makers, e.g., glutamate; others are helix-breakers, e.g., proline, glycine, and serine (16). Also, because of the orientation of peptide bonds in their backbones, α-helices have an intrinsic dipole (17–19), equivalent to a charge of about $+\frac{1}{2}$ e at the NH_2-terminus and $-\frac{1}{2}$ e at the COOH-terminus (e = elementary charge). The dipole exists even when the α-helix is part of a longer peptide. Negatively charged residues (Asp/Glu) at the NH_2-terminus interact favorably with the dipole, as do positively charged residues (Arg/His/Lys) at the COOH-terminus. These interactions can help to stabilize an α-helix and, in fact, many α-helices in native protein have these residues in the appropriate position (18). α-Helicity, if present, can have many implications for the composition of antigenic peptides.

The extended (i.e., β) peptide conformation is common in native proteins and can also be amphipathic. Unlike α-amphipathicity, β-amphipathicity is not yet implicated in T cell antigenicity. We shall use the word *β-propensity* to connote a tendency to β-conformation. Similarly, peptides with a tendency to α-helicity have α-propensity. The terms "turn propensity" and "coil propensity" are self-explanatory.

Confirmation of the correlation of amphipathicities, propensities, and other properties with immunodominance requires a statistical test. Classical statistical methods are inappropriate for protein analysis because they require analytic description of the parent distribution (is the distribution normal? chi-squared?, etc.). The "Materials and Methods" provides a novel and appropriate statistical test for significance in protein (or DNA) data bases, made practicable by Monte Carlo computer experiments. This test can confirm the correlation between a property and peptide antigenicity. Significant statistics will be used in predictive schemes elsewhere (Margalit et al., manuscript in preparation).

METHODS

1. Antigenic data base. Table I lists the antigenic sites as this paper uses them in statistical tests. The selection criteria for this particular list are: the sites *a*) were reported to be immunodominant in the response to a protein, *b*) were known to the authors prior to February 21, 1986, and *c*) are less than 21 residues long. The restrictions involve arbitrary cut-offs, but were necessary *a*) to close the statistical data base and *b*) to localize immunodominant sites. (Antigenic sites much longer than 21 residues probably do not localize their immunodominant site sufficiently.) The entries in Table I are, for each experiment, representative of the shortest peptide capable of near-maximal T cell stimulation. Such peptides are usually obvious from the experimental data: deletion of critical residues generally produces a precipitous drop in antigenic activity. When the experiments did not localize the end residues of an antigenic site, the criteria given in Table I were applied to give a definite peptide suitable for statistical testing. In the absence of a registry of immunodominant sites, these criteria were as objective as possible.

Thus, Table I does not give the antigenic sites as reported in the literature, but, within the context of statistical examination, is as faithful as possible to experimental results. In particular, for reasons detailed below, in the absence of explicit information, it was safer to treat certain peptides as though they had been produced by tryptic or cyanogen bromide cleavage. For technical reasons explained under "α-amphipathicity" (subsection 3 of *Methods*), the antigenic sites sperm whale myoglobin 69–78 and influenza hemagglutinin 111–119 were extended to length 11 when α- and β-amphipathicity were examined. Table I is *not* a summary of the relevant experiments, but is rather a means of reproducing our statistical results.

2. Statistical methods: site statistics. Assume that any site (i.e., any peptide, not necessarily antigenic) within a protein can be associated with a number, a site statistic A. The nature of this number need not concern us yet; it may reflect α-propensity, sequential amphipathicity, the absence of prolines within the site, etc. Our problem is to determine the significance of the site statistic A as a correlate of immunodominance.

Overall statistics. The mechanics of this determination are as follows: there are 12 proteins in the data base and a total of 23 antigenic sites. *a*) Calculate the site statistics A_0 for the antigenic sites and then *b*) sum these to produce an overall statistic S_0 for the antigenic sites.

Statistical significance and anti-significance. We now wish to assign a significance to S_0. As an example of (one-tailed) statistical significance, imagine a normal (Gaussian) curve. The significance of a number S_0 is the area to the right of S_0 under the curve. This is the probability of drawing a random number S exceeding S_0 from that normal population.

[1] Abbreviation used in this paper: APC, antigen-presenting cell.

TABLE I

Antigenic sites as used in the Monte Carlo computer experiments[a]

Status of Residues near the Peptide COOH-Terminus		
* = Cleavage Restriction. COOH-terminal Arginine or Lysine		
# = Cleavage Restriction. COOH-terminal Methionine		
K = A Lysine Known to be Necessary for Peptide Antigenicity		
Sperm Whale Myoglobin		
68–78	V[b] L T A L G A I L K K	(20)
102–118	K Y L E F I S E A I I H V L H S R	(21)
132–146	N K A L E L F R K D I A A K[c] Y	(22)
Pigeon Cytochrome *c*		
94–104	L I A Y L K[d] Q A T A K[d]	(12)
Influenza Hemagglutinin Protein A/PR8/48		
109–119	S[b] S[b] F E R F E I F P K	(24)
129–140[e,f]	N G V T A A C S H E G K	(25)
302–313[g,e,f]	C P K Y V R S A K L R M	(25)
Pig Pro-Insulin[h]		
A4–14[i,j]	E Q C C T S I C S L Y	(4)
B5–16	H L C G S H L V E A L Y	(27)
Hen Lysozyme		
46–61[e,k]	N T D G S T D Y G I L Q I N S R	(28)
74–86	N L C N I P C S A L L S S	(29)
81–96[e,k]	S A L L S S D I T A S V N C A K	(29)
108–119[j]	W V A W R N R C K G T D	(30)
Hen Ovalbumin		
323–339[e,k]	I S Q A V H A A H A E I N E A G R	(31)
Hepatitis Pre-S		
120–132	M Q W N S T T F H Q T L Q	(32)
Foot and Mouth Disease Virus VP1		
141–160	V P N L R G D L Q V L A Q K V A R T L P	(33)
Beef Cytochrome *c*		
11–25	V Q K C A Q C H T V E K G G K[m]	(34)
66–80[e,f]	E Y L E N P K K Y I P G T K M	(10)
Hepatitis B S-Antigen HBsAg/adw		
38–52	S L N F L G G T T V C L G Q N	(35)
95–109	L V L L D Y Q G M L P V C P L	(35)
140–154	T K P S D G N C T C I P I P S	(35)
Lambda Repressor		
12–26	Q L E D A R R L K A I Y E K K[n]	(36)
Rabies Spike Glycoprotein		
32–44[e,m]	D E G C T N L S G F S Y M	(37)

[a] Each amino acid is represented by its single-letter code which is the first letter of its name, except for: arginine (R), aspartic acid (D), asparagine (N), glutamine (Q), glutamic acid (E), lysine (K), phenylalanine (F), tryptophan (W), and tyrosine (Y).

The primary source for each antigenic sequence is indicated to its right. The numbering of the antigenic sequence, when available, was taken from the primary source, as was the parent protein sequence used in the Monte Carlo computer experiments. When otherwise unavailable, the protein sequences were taken from the National Biomedical Research Foundation protein sequences database (Georgetown University, Washington, D.C.). If the termini of an antigenic peptide were not determined by experiment, the peptide was taken to be of length 11 or 12 and centered around known critical residues. Length 11 was used if the number of critical residues was odd, 12 if it was even. If the original paper contained insufficient information to eliminate biasing of the COOH-terminus residue by either tryptic or cyanogen bromide cleavage, cleavage restriction was imposed in the statistical analysis. The *Methods* section explains these issues more fully.

[b] NH_2-extensions of the antigenic site required by amphipathicity statistics; these residues are not part of the minimal stimulating peptide. Peptides containing them retain antigenicity, however, according to the primary source. See under "α-amphipathicity" in subsection 3 of *Methods* for further explanation.

[c] Lysines necessary to antigenic activity, according to primary source.

[d] Lysines necessary to antigenic activity (23).

[e] Substitutions of N-129, S-136, and E-138 established criticality of those residues, and the peptide shown overlapped with an antigenic "cleavage fragment" in the primary source. As explained in *Methods*, the most conservative course is to represent the antigenic site by the peptide shown.

[f] Insufficient information to ascertain non-biasing of COOH-terminal residue.

[g] A homologous site in H3 influenza hemagglutinin protein is recognized by human T cells (26).

[h] Pro-insulin was used as the parent protein: A- and B-chains were thus combined into a single protein.

[i] Termini not determined by experiment.

[j] A8–10 are critical residues (4).

[k] Tryptic cleavage, according to primary source.

[l] 113–114 are critical residues (30).

[m] Residues 23–25 are thought to be necessary to antigenic activity, although this is not yet established (34).

[n] Cyanogen bromide cleavage, according to primary source.

A property is unusually frequent in the antigenic sites if it has a low significance (p value). Likewise, it is unusually infrequent if it has a high p value. We call this anti-significance (this term is not standard statistical terminology but is convenient for our purposes). Anti-significance indicates statistical significance of an anti-variate (e.g., presence of a residue instead of its absence) and, like significance, it signals an unusual occurrence. If p is the variate significance, then the significance of the anti-variate is (1 − p): a p of 0.95 is as remarkable as a p of 0.05. This is just one-tailed statistical significance for the left tail of a distribution, rather than its right tail.

To assign a significance to the antigenic statistic S_0, we must examine what would happen if the antigenic sites were chosen from a random population. For random sites, we can follow the procedure above: *a*) compute the site statistics A and then *b*) add the site statistics together to obtain an overall statistic S. The probability that S exceeds S_0 is the statistical significance of S_0. Our main difficulty lies in defining a "random site."

Definition of "random." We give an exact definition of "random" below. The following metaphor is useful: imagine that the protein sequences are listed in parallel lines on the ground. Sticks of the appropriate lengths underscore the antigenic sites within the sequences. We shall have a computer pick up the sticks one by one and throw them at random back down onto the sequence from which they came. Each stick underscores a peptide: these peptides are our new "random sites." Repeating this procedure generates more random sites. For reasons given below, sometimes we shall insist that certain random sites must have a specific amino acid at their COOH-terminus. This does not change the mechanics of the random selection; it only restricts where the corresponding stick is allowed to fall.

Before an experiment is performed, the positions of the antigenic sites are unknown. The above procedure is therefore a reasonably faithful representation of possible outcomes. Note that we never scramble the protein sequence in any way. This common (and commonly fallacious) practice is inappropriate here, because scrambled proteins do not represent possible experimental outcomes.

We now define "random" precisely: consider, for example, beef cytochrome *c*. It contains two antigenic sites, both of length 15. One of these was produced by cyanogen bromide reaction and had to end in a methionine. We choose two "random" sites of length 15 within beef cytochrome: for one of these random sites, any site of length 15 is equally probable; for the other, any site of length 15 ending in a methionine is equally probable (so the second random site is chosen from a much more restricted class of sites). Choose other random sites within the other proteins, according to the corresponding lengths and number of the antigenic sites within each protein. If an antigenic site was produced by a tryptic digest, then the corresponding random site should end in arginine or lysine; likewise, if it was due to a cyanogen bromide reaction, then the corresponding random site should end in a methionine. We call this cleavage restriction.

Why enforce a cleavage restriction? Tryptic digests (which force the terminal residue of an antigenic site to be either arginine or lysine) systematically bias the COOH-terminal residue of an antigenic site. Lysines at the COOH-terminus of antigenic sites turn out to be important. Cleavage restriction controls the bias that tryptic digests and cyanogen bromide reactions introduce into the COOH-terminal residues.

In theory, all possible experimental biases (even including the experimenter's avoidance of certain amino acids in peptide synthesis) should be included in the definition of "random." In practice, this is not possible. It is even undesirable if it restricts the pool of possible random sites too much. For example, in the extreme case, let a particular "random" site always be chosen to coincide with the corresponding antigenic site (i.e., in the stick-throwing analogy, the stick underlining the antigenic site never moves). The contribution A_0 that the site makes to the statistic S is fixed. This is equivalent to eliminating the antigenic site from the data base.

The emphasis is therefore more on preventing a statistical bias than on imitating a physical process. Cleavage restriction does not imitate proteolytic cleavage perfectly. For example, a tryptic digest is unlikely to produce the hypothetical site Ala-Leu-Val-Gly-*Lys*-*Lys*-Thr-Tyr-Cys-*Lys* because of the presence of the two internal lysines. Likewise, a tryptic fragment follows a lysine or arginine in the original protein sequence. Similar considerations hold for cyanogen bromide. The definition of "random" does not encompass these and similar facts. In view of the preceding paragraph, however, and especially because these facts should not bias our statistics, they have been ignored.

In practice, the information required to eliminate experimental biases is not always available. In the absence of the requisite information, a site was always assumed to be subject to bias. The best example of this is the antigenic site influenza hemagglutinin 129–140 in Table I (25). This site was localized by examining the antige-

nicity of hemagglutinin variants and a "cleavage peptide" (the cleavage method and the precise peptide were unspecified in the reference). The most conservative course is to assume that the cleavage localizing the antigenic site was tryptic, and then to subject the site to cleavage restriction.

Our site statistics are also systematically influenced by site length, so control of site length is essential. Because the "random" selection of potential antigenic sites incorporates controls for site length and cleavage methods, these two variables should not bias our statistics, in particular our COOH-terminal statistics.

Residue restriction. Unless otherwise stated, cleavage restriction is always used to control the COOH-terminus of the random sites. The one exception, used in special cases only, is residue restriction. Here the antigenic sites are classified by their COOH-terminal residue: Arg, Lys, Met, and other. Random sites are chosen only from the same class as the corresponding antigenic site. COOH-terminal lysines will turn out to be significant correlates of antigenicity: the intent of residue restriction is to remove the effects of COOH-termination in lysine and measure independent effects from other sources. By including restrictions on arginine and methionine, residue restriction continues to prevent bias from cleavage methods.

A practical example best illustrates the reason we use residue restriction in certain analyses. Preliminary statistics showed that random sites having lysine at their COOH-terminus have a higher α-amphipathicity (as quantitated below) than random sites with other COOH-terminal residues. They also showed that COOH-terminal lysines were unusually frequent among the antigenic sites. Because lysine is a very hydrophilic amino acid, the COOH-terminal lysines alone may be causing a high antigenic α-amphipathicity. Because residue restriction controls the number of sites with COOH-terminal lysines, any significance of α-amphipathicity under residue restriction argues that α-amphipathicity contributes to antigenicity independently of the COOH-terminal lysines. In general, if a statistic retains its significance under *Residue Restriction*, its significance cannot be due to the unusual frequency of COOH-terminal lysines in Table I.

Monte Carlo computer experiments. We still need a way to estimate the probability that S exceeds S_0. This can be done by a computer employing Monte Carlo computer experiments (38). The computer chooses random sites a large number of times. Each time, the "random" overall statistic S is computed and compared to S_0. The proportion of times that S is greater than or equal to S_0 is the required estimate of the statistical significance of S_0. The more times the computer choose random sites, the better this estimate of significance. Each event ($S_0 \leq S$) is one binomial trial, a 1 or a 0, and an appeal to the binomial distribution (39) shows that 50,000 computer trials given an estimate of significance accurate to about ±0.005. Accordingly, this was the number of trials used.

The following analogy justifies the whole procedure. Suppose you have a set of loaded dice (dice = proteins) which gives sixes more frequently than they should (faces = antigenic sites, numbers on the faces = measure of α-amphipathicity, e.g.). You throw all the dice, some perhaps more than once (find some antigenic sites, perhaps more than one per protein), and total the resulting numbers (total the α-amphipathicity scores A for the antigenic sites found). This gives a total S_0 (overall statistic) for the loaded dice.

You now take a set of dice which are not loaded (the computer) and throw them several times. If their total S is consistently less than the total S_0 from the loaded dice, then the original dice must indeed have been loaded (e.g., the antigenic sites tend to be more α-amphipathic than they should be on the basis of chance alone). The total S is statistically more important than the individual numbers A on the dice because any individual die is subject to considerable random fluctuation. Despite this individual fluctuation, a systematic bias will show itself in the total S_0 on the original dice.

This method will determine whether the antigenic sites are "loaded" with respect to certain peptide properties.

3. Statistics representing the properties: block averaging and maximization. The site statistics chosen to represent the properties are to some extent arbitrary. To facilitate programming, many of these statistics are generated from more elementary *block statistics*, numbers that are attached to peptides of a fixed length (blocks) within the protein. The block statistics must then be converted into site statistics. There are at least two reasonable procedures for doing this: *a*) "block averaging" and *b*) "block maximization" block averaging means averaging the block statistic over all the blocks completely contained within the antigenic site (similarly, block maximization). If an antigenic site contains many "ordinary" blocks along with some immunodominant blocks, averaging dilutes the contribution that the immunodominant blocks make to the site statistic. Hence, block maximization is usually the procedure of choice.

The statistics are indexed (A, C.i, etc.) identically in the following and in Table II. For reasons detailed elsewhere (Margalit et al., manuscript in preparation, and Cornette et al., manuscript in preparation), this paper uses the Fauchere-Pliska (40) scale as a measure of amino acid hydrophobicity.

TABLE II

Statistical significances, cleavage restriction

Statistic	Significance p	
	p	(1 − p)
A. α-Amphipathicity	0.017	
B. β-Amphipathicity	0.855	
C. α-Helical Properties		
i. α-Propensity	0.031	
ii. Residues (Helix-Makers and -Breakers)		
a. Glutamate Presence	0.627	
b. Proline Absence	0.098	
c. Glycine Absence	0.048	
d. Serine Absence	0.683	
iii. Moment (Helical Dipole)		
a. Charge	0.095	
b. Lysine Charge	0.042	
c. Histidine Charge	0.096	
d. Arginine Charge	0.713	
e. Aspartate Charge	0.165	
f. Glutamate Charge	0.524	
iv. COOH-terminal Lysines		
a. 1-Ultimate Lysine	0.005	
b. 2-Ultimate Lysine	0.010	
D. β-Propensity	0.152	
E. Turn Propensity	0.656	
F. Coil Propensity	0.976	(0.024)
G. Segmental amphipathicity		
i. Differential Hydrophobicity:	0.843	
ii. Maximum Differential Hydrophobicity:	0.887	

A. α-Amphipathicity. The first property to be examined is α-amphipathicity. The intensity of the discrete Fourier transform provides a site statistic (7). The Fourier transform picks out periodicities in a sequence of numbers: in this case, it can pick out the 100° periodicity of hyrophobicities corresponding to an amphipathic α-helix. a) Divide the proteins into overlapping blocks of length ℓ. The first block extends from residue 1 to residue ℓ, the second block from residue 2 to residue $\ell + 1$, etc. (If the protein has length L, then the number of blocks is $L - \ell + 1$ (e.g., a protein of length ℓ contains exactly 1 block).) $\ell = 11$ is appropriate, because Fourier transforms with smaller ℓ do not always reflect periodicities faithfully (Cornette et al., manuscript in preparation). Because two minimal antigenic sites, sperm whale myoglobin 69–78 and influenza hemagglutinin 111–119, in Table I, are of lengths less than 11, these sites are extended *for amphipathicity statistics only* to make their lengths 11. The resulting peptides retain near-maximal antigenicity (20, 24). The NH_2-terminus rather than the COOH-terminus was extended because, as is shown later, COOH-terminal lysines correlate with antigenicity.

Let h_k be the hydrophobicity of the k^{th} residue in the protein and $\bar{h}_k$ be the average hydrophobicity of the k^{th} block (which consists of residues k to $k + \ell - 1$). The intensity of the discrete Fourier transform of the residue hydrophobicities is:

$$I(k,\theta) = \{\left|\sum_{j=k}^{k+\ell-1} (h_j - \bar{h}_k)\sin(2\pi\theta j/360)\right|^2 + \left|\sum_{j=k}^{k+\ell-1} (h_j - \bar{h}_k)\cos(2\pi\theta j/360)\right|^2\}^{1/2} \quad (1)$$

The Fourier intensity can again be converted to a site statistic in many different ways. The maximal α-intensity is an appropriate choice: b) For each block, take the maximum of the Fourier intensities at $\theta = 80°, 85°, 90°, \ldots, 120°$. (Unlike the counterpart statistic (7), the maximal α-intensity does not depend on values outside the 80° to 120° range.) Because the Fourier intensity at 100° corresponds to the amphipathicity of residues in an exact α-helical conformation, the maximization around 100° producing the maximal α-intensity allows for deviation from exact α-helicity. This maximization provides a block statistic which is then block-maximized (as described at the state of subsection 3) to yield a site statistic. Because maximal α-intensity is the only statistic we use to represent α-amphipathicity, we shall refer to it as "α-amphipathicity" (see *A* above).

B. β-Amphipathicity. We define the maximal β-intensity similarly, but maximize the Fourier intensities at $\theta = 160°, 165°, 170°, 175°$, and 180°. (By symmetry, the intensities between 180° and 200° are irrelevant (Cornette et al., manuscript in preparation).) 180° corresponds to an exact β-sheet. As in "α-Amphipathicity" above, a block length $\ell = 11$ was used and the two shortest antigenic sites

are again extended to this block length.

C. *α-Helical properties.* The α-helical conformation is well-investigated, and as such has many different measures and implications. The number of statistics presented reflect this depth.

i. *α-Propensity.* *a*) Divide the proteins into overlapping blocks of length ℓ (we take $\ell = 9$ because this is the length of the shortest antigenic site). *b*) Sum the appropriate values in Table I of Garnier et al. (41) (which we refer to as G-O-R Table I) to calculate the directional α-helical information for the central (5th) residue in the block. This generates a block statistic. In a departure from the usual procedure, block average to produce a site statistic. This gives the tendency of the entire site to form an α-helix. (Block maximization would reflect the residue most likely to be in α-helical conformation: if isolated, this residue is probably not very important.) Because this statistic attempts a complete representation of α-propensity, we shall refer to it as α-propensity. Note that the G-O-R-Tables are based on the statistics of native proteins (41), not short peptides. This distinction will turn out to be important.

ii. *Residue presence and absence.* Some residues, notably glutamate, are "helix-makers" whereas others, notably proline, glycine, and serine, are "helix-breakers". Helix-makers are frequently found in α-helices, helix-breakers infrequently. The following statistic, residue presence, tests whether a residue occurs more frequently in antigenic sites than at random. *a*) Assign the residue in question a value of 1 and all other residues a value of 0. *b*) Average these numbers over each site to produce a site statistic and add the site statistics together in the usual way to produce an overall statistic. Presence of the residue in question increases this statistic. Changing the sign of the residue values yields residue absence, which reflects the absence of the residue in question.

iii. *The moment.* This is defined in conjunction with a set of amino acid values. The values are numbers which are assigned to the amino acids, e.g., hydrophobicity, charge, etc. Unusual moments reflect nonrandom distribution of the values along the length of a site. We shall be most interested in charge moments. *a*) Divide the protein up into overlapping blocks of length ℓ We use $\ell = 9$. *b*) Assign to all the residues in a block numbers indicating their signed distance from the center of the block. If ℓ is odd, the center residue gets a 0, the carboxy-terminus residues are labeled 1, 2, 3, . . . in sequence from the center, while the amino-terminus residues are labeled −1, −2, −3, ... If ℓ is even, there is no center residue, but by analogy with the above, the residues next to the center are labeled ½ and −½, the ones next to those 3/2 and −3/2, and so forth. *c*) Multiply the numbers by the value of the amino acid occupying the position. *d*) Add the resulting products together. This is the moment of the values within the block. Maximizing this block statistic produces a site statistic.

The moment of charge is large whenever either negative side-chains (Asp/Glu) are near the NH_2-terminus or positive side-chains (Arg/His/Lys) are near the COOH-terminus. This nonrandom charge distribution is the one required for favorable interaction with the α-helical dipole and would be expected to correlate with α-helices.

We examine the moments corresponding to the following amino acid values: a) Charge: Arg = Lys = 1, His = 0.3 (His is somewhat arbitrary), Asp = Glu = −1, all others = 0; b) lysine charge: Lys = 1, all others = 0; and c) aspartate charge: Asp = −1, all others = 0. arginine, histidine, and glutamate charges are defined analogously.

iv. *COOH-terminal lysines.* The following are 1/0 statistics, i.e., statistics which take the value 1 if the site has a certain property and 0 otherwise. The 1-ultimate lysine is defined as follows: if the end-residue on a site is a lysine, then the site statistics is a 1. Otherwise the site receives a 0. The 2-ultimate lysine is similarly defined: the site receives 1 if there is a lysine in either of the last two positions and 0 otherwise. (None of the antigenic sites in Table I has an antepenultimate lysine, so we arbitrarily terminate the series of ultimate lysines at 2.) The overall statistic S corresponding to the 1-ultimate lysine is the sum of the site statistics and is just the number of sites having lysine at their COOH-terminus. A similar relationship holds for the other ultimate lysines.

The three statistics represent β conformations, turns, and coils.

D. *β-Propensity.* This is exactly analogous to α-propensity except that we use Table 2 of Garnier et al. (41). Because it is the only attempt to represent β-propensity, we shall refer to it as β-propensity.

E. *Turn propensity.* This is analogous to α-propensity, except that we use Table 3 of Garnier et al. (41).

F. *Coil propensity.* This is also analogous to α-propensity, except that we use Table 4 of Garnier et al. (41).

G. *Segmental amphipathicity.* We give two site statistics that can represent segmental amphipathicity: many more are possible.

i. *Differential hydrophobicity* (7). *a*) Divide the proteins into overlapping blocks of length 2ℓ (we take $\ell = 4$). *b*) For each block, take the sum of the hydrophobicities of the ℓ residues on either end of the block. c) Take the absolute value of the difference of the two sums. This yields a block statistic; block maximization as described above yields a site statistic.

ii. *Maximal differential hydrophobicity.* *a*) Divide the proteins into overlapping blocks of length ℓ (we take $\ell = 4$). *b*) For each block, take the sum of the hydrophobicities of the ℓ residues. *c*) For every pair of non-overlapping blocks within the antigenic site, find the absolute value of the difference of block sums. *d*) The site statistic is the maximum of these differences. Maximum differential hydrophobicity systematizes the procedure that Corradin et al. (10) carried out by eye for eight antigenic sites.

4. *Correlations.* For any pair of site statistics X and Y, and for any 23 sites (whether random or antigenic), we calculate $r = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y)$, the correlation coefficient of the 23 (X,Y) pairs. ($r = 1$ for perfect correlation (e.g., X = Y), $r = -1$ for perfect anti-correlation (e.g., X = −Y), and $r = 0$ if X and Y are independent.) r is itself an overall statistic, and its expectation $\bar{r}$ reflects the coupling of X and Y in random sites. Denote the r for the antigenic sites by r_0. r_0 has a statistical significance which can be estimated by Monte Carlo computer experiments. Because r_0 reflects the coupling of X and Y over the actual antigenic sites, a statistically significant r_0 may reflect an (X, Y) pair which is unusually coupled within the antigenic sites.

We can illustrate the practical use of correlations by an example. Take two site statistics (A) called X and Y (e.g., α-amphipathicity and α-helicity). X and Y may be correlated in random sites (e.g., many α-helices in native proteins are amphipathic. X and Y are both statistically significant: does the significance of X depend on its correlation with Y? If so, r_0 should not be statistically significant: X and Y should be no more tightly coupled in the antigenic sites than they are in random sites. However, if the correlation of X and Y in the antigenic sites is statistically significant compared to their correlation in random sites, X and Y are more tightly coupled in antigenic sites than they would be at random. Hence, one might infer that X and Y contribute independently to antigenicity. When X is α-amphipathicity, the two shortest antigenic sites in Table I must again be extended to length 11 as described in subsection 3 under "α-amphipathicity".

RESULTS AND DISCUSSION

A. Experimental findings. Tables II and III were obtained under cleavage restriction. Because results for residue restriction (not shown) were similar, we conclude that other significances were independent of COOH-terminal lysines.

The results in Tables II and III, detailed in brief below and discussed in greater depth in Part C, have important implications for the manufacture of peptide vaccines. These implications are as follows: if possible, peptides vaccines should probably be those protein segments *a*) which have a propensity to form amphipathic α-helices, *b*) which do not have regions with a propensity to coil conformations, and *c*) which have a lysine at their COOH-terminus. The last two observations are of particular use in manufacturing peptides vaccines: they indicate where the synthetic peptides should be terminated.

α-Helical properties. All of these were strongly represented in the antigenic sites, suggesting that many antigenic sites take an α-conformation. Of these properties, α-amphipathicity was the most significant. The correla-

TABLE III

Correlation statistical significances, cleavage restriction

Statistics X and Y	Expected Correlation $\bar{r}$	Antigenic Correlation r_0	Significance p
X. α-Amphipathicity Y. α-Propensity	0.260	0.479	0.136
X. α-Propensity Y. β-Propensity	−0.368	−0.652	0.954
X. β-Propensity Y. Turn Propensity	0.082	0.452	0.041
X. β-Propensity Y. Coil Propensity	−0.022	0.445	0.013

tion of α-amphipathicity and α-propensity had a significance of $p = 0.136$, suggesting that the two properties may make independent contributions to T cell antigenicity.

COOH-terminal lysines. Lysines were unusually frequent at the COOH-termini of antigenic sites. This could not be an artifact of tryptic digestion, because cleavage restriction controlled potential biases from that source. As Table I records, experimental removal or substitution of these COOH-terminal lysines often destroys antigenic activity. This fact, never before noted as a general observation, may be useful in designing the COOH-termini of synthetic peptides used for vaccination. In addition, the significance of α-properties and COOH-terminal lysines may suggest something general about the chemistry of T cell recognition (see Part C).

Conformational propensities. The sites displayed some β-propensity, but no β-amphipathicity. Turn propensity was not significant, but coil propensity was in fact anti-significant, thus coils were notably absent in the antigenic sites. β-Propensity was significantly correlated with turn and coil propensities and significantly and strongly anti-correlated with α-propensity. These significances perhaps suggest that some antigenic sites take β-conformations but have their β-propensity masked by th anti-correlating α-helices.

Segmental amphipathicity. Segmental amphipathicity was not statistically significant. In case some sites were masking the segmental amphipathicity of others, we tested several different subsets of the antigenic sites, in particular, those in hen lysozyme for which segmental amphipathicity was first invoked (7, 13) and the subset in Table III of Corradin et al. (10). No subset tested showed significant segmental amphipathicity.

Section C (*Discussion of Experimental Results*) offers explanations for these findings. Before this, however, we give a discussion of our statistical methodology.

B. Discussion of the statistical approach: equidistribution of ignorance. Our statistics are based on the "equidistribution of ignorance." Because we have no particular reason to favor one of several alternatives, we assign all the alternatives an equal probability. In some branches of science, in particular statistical physics, equidistribution of ignorance can be justified from first principles by the so-called ergodic theorems (42). No such justification can be invoked in biology. In biology, the equidistribution is purely the statement of our own ignorance.

In this report, our initial ignorance is tempered only by knowledge f experimental conditions, such as the use of tryptic digestion or cyanogen bromide reaction. Our definiti n of "random" includes these factors. Our admission of ignorance, the definition of "random," provides a benchmark against which the Monte Carlo method is applied, and the value of a statistic S_0 is compared against the values S we expect. Any statistic S which is statistically significant can then be used predictively. This process reflects a change in the state of our ignorance.

Stratification. In theory, we can use our new knowledge to change the *a priori* probability of various sites. Take α-propensity as an example. Table II indicates that that experiment tends to find sites with a larger α-propensity than the random selection produces ($p = 0.031$). Our definition of "random" can be altered to encompass this fact. We can stratify the random sites into several groups according to their α-propensity. The computer experiment can be run again, but this time we always select a particular random site from the same stratum as the corresponding antigenic site. (This stratification is analogous to the stratification by cholesterol level that would determine whether smoking, independent of cholesterol level, was a factor in heart disease.) Residue restriction is a particularly simple form of stratification. The sites are stratified into four classes on the basis of their terminal residues: Arg, Lys, Met, and other. Random sites are then chosen from the same stratum as the corresponding antigenic site.

Stratification could, for example, separate the relative contributions of α-propensity and α-amphipathicity to antigenicity. We attempted the separation using correlation significance instead because stratification requires more programming effort.

Avoiding spurious significance. This paper obeyed one very important methodological maxim: we tested only those statistics consistent with a physical hypothesis. If, without a reason to do so, we had tested all 20 amino acids for some property, one amino acid would probably have been significant at $p < 0.05$ by chance alone. Approaches not based on physical theories run a higher risk of producing spurious significances.

Physical interpretation and predictive use of statistics. Despite our emphasis on physical interpretation, the G-O-R Tables (41) predict native structure with only 57% accuracy (43). It is unlikely that our antigenic propensities correlate much better with either native or peptide conformations. How then should the statistics be interpreted?

The significances of various conformational propensities depend on the statistics used to measure them. If, for example, the G-O-R Tables are applied to a residue without considering the surrounding residues, there is no statistical significance. The second law of thermodynamics (as applied to information) states that every irreversible transformation or simplification of raw data destroys information, e.g., if someone else cannot recover your data base after you have manipulated it, you have destroyed information. Our statistical method, because it can utilize raw data, demonstrates this destruction of information as a loss of statistical significance.

By the same token, if a statistic is significant, we can be confident that every step in its production probably preserved some information, especially if the statistics were based on a physical hypothesis. The G-O-R Tables can be considered statistical correlates of the free energies of amino acid interactions. Presumably, enough information about these free energies was preserved to make our statistics significant. Our statistics will not always yield correct predictions, conformational or otherwise, however. In fact, the conformation f a peptide recognized by a T cell may perhaps sometimes differ from its conformation in native protein (7, 44), making conformational verification even more difficult.

In the terminology of the dice-rolling analogy in the *Statistical Methods*, if some dice are loaded towards sixes, we can detect this by rolling the dice a sufficient number of times. Despite being loaded, the dice will sometimes produce ones. In the case of the T cell antigenic sites, the sites are loaded with, for example, extra pro-

pensity towards α-conformations. In analogy with the dice, the statistical methods of this study can only give rise to statistical predictions, and sometimes the conformational predictions from propensities will be wrong. Despite this, one is better off predicting from statistically significant parameters than at random.

These considerations should not obscure the fact that the statistical significances reflect *biochemical* properties of T cell antigenic sites. This paper used p values as an instrument to discover biochemical properties of T cell antigenic sites and evaluate the extent to which these properties are significant indicators of T cell antigenicity. Later reports[2] will use the statistics presented here predictively.

C. Discussion of experimental results. Tables II and III summarize our results.

Our statistics are consistent with a "conformational hypothesis": helper T cell immunodominant sites tend to be peptides *with strong conformational propensities* that can stabilize under hydrophobic interaction with class II MHC proteins. The conformational hypothesis is an extension of the amphipathicity hypothesis (7) which does *not* consider conformational propensities.

α-Properties. A consistent significance for α-properties emerged, suggesting that most T cell antigenic sites take an α-helical conformation. α-amphipathicity and α-propensity are both significant ($p = 0.017$ and $p = 0.031$). Moreover, their correlation may also be significant ($p = 0.136$). Hence, α-amphipathicity may be a significant factor in T cell antigenicity independent of its correlation with α-propensity. Antigens stimulating helper T cells may bind to the class II protein through hydrophobic interaction (1, 13, 15); because recognition occurs at the interface between a class II protein at the antigen-presenting cell surface and an aqueous environment, α-helical amphipathicity may help to stabilize antigens in α-helical conformation. This forms the basis of the so-called amphipathicity hypothesis (7).

Helix-makers and -breakers. α-Helical conformations, whether amphipathic or not, should display the characteristics mentioned in the *Introduction.* The helix-breakers proline and glycine should be infrequent ($p = 0.098$ and $p = 0.048$). The next helix-breaker tested, serine, was not statistically significant ($p = 0.683$). Similarly, the helix-maker glutamate was not present in unusual amounts ($p = 0.627$). In accord with the end of the *Discussion on Statistical Methods*, tests for helix-making and -breaking significance ended here.

The intrinsic dipole. The next physical consequence of α-helicity we examined was the intrinsic dipole and its favorable charge distribution, represented by the moment of amino acid charge. The moment of amino acid charge has a significance of $p = 0.095$. (Recall that our statistics were controlled for tryptic digestion and its biasing of charge distribution. Those biases could not influence the results.) When the moments for individual residue charges are examined, lysine and perhaps histidine are significant ($p = 0.042$ and $p = 0.096$), whereas arginine, aspartate, and glutamate are not ($p = 0.713$, 0.165, and 0.524, respectively).

COOH-terminal lysines. Once attention is drawn to lysine, Table I shows its positional preferences quite strikingly. Lysine, appearing near the COOH-terminus of antigenic sites far more often than its frequency in proteins warrants, is often necessary for antigenic activity. The significance of the 1- and 2-ultimate lysines in Table II is remarkable ($p = 0.005$ and $p = 0.010$).

The ultimate lysines contain a subtle point. Because the 1- and 2-ultimate lysines are more statistically significant than α-amphipathicity ($p = 0.017$) and α-propensity ($p = 0.031$), these last two qualities cannot convincingly explain the tendency of antigenic sites to terminate in lysines.

Recall that the directional information of the G-O-R Tables (41), the basis of our "propensity" statistics, are based on native peptide conformations (45). The G-O-R Table I, if examined, does indeed reflect the stabilizing influence of COOH-terminal lysines on native α-helices (about 6 kcal/mol = 10 kT, for 10 residue helices in native protein (17)). Despite this, relative to α-propensity, COOH-terminal lysines are unusually significant. This suggests that COOH-terminal lysines stabilize α-helices much more in free peptides than they do in native proteins.

In native proteins, α-helices have no free backbone charges. Free peptides, by contrast, have an extra free charge on both their NH_2- and COOH-termini. If a free peptide has an α-helical conformation, electrostatic interactions overwhelmingly favor the conformation placing the terminal carboxy-charge away from the backbone carbonyl groups near the α-helical axis. This swings the side-chain underneath the helical dipole in the free peptide, rather than to one side as occurs in the native protein. (Similar considerations apply to the NH_2-terminal.) The dipole field is stronger along the dipole axis, so terminal-charged side-chains interact more strongly with the electrostatic field of the intrinsic dipole in free peptides than in native proteins. Hence, favorable charge distributions are much more decisive in stabilizing α-helices in free peptides than they are in native proteins. This effect may be most apparent with lysine because, of all charged residues, lysine has both the most mobile side-chain and the most localized charge.

For very short peptides which cannot have many α-helical hydrogen bonds to stabilize them, the extra stability may be very important. For example, residues 111–119 in influenza hemagglutinin are consistent with this hypothesis, as are residues 69–78 in sperm whale myoglobin, the latter ending in not just one, but *two* lysines. The penultimate lysine probably further stabilizes the conformation by electrostatic interaction with the peptide carboxy-charge.

Other explanations for the COOH-terminal lysines are possible. Although they might result from enzymatic cleavage during antigen processing, there is little evidence for marked enzymatic specificity in antigen-processing cells (46). Another plausible explanation might have some or all of these lysines interacting directly with T cell receptors or class II molecules on APC. Pincus et al. (11), for example, implicate α-helical conformations in the antigenic activity of pigeon cytochrome c residues 94–104. They could not distinguish whether the COOH-terminal Lys-104 was necessary for T cell receptor interactions or for maintenance of an α-helix, however. Indeed, if the Lys-104 depends on the α-helical conformation for contact with the T cell receptor, distinguishing

[2] Margalit, H., J. L. Spouge, J. L. Cornette, K. Cease, C. DeLisi, and J. A. Berzofsky. 1986. Prediction of immunodominant helper T-cell antigenic sites from the primary sequence.

these alternatives could be meaningless. Possibly, the distinction could be drawn in other cases, and some COOH-terminal lysine may be necessary for T cell interaction alone.

Other conformational properties. The significance of β-propensity ($p = 0.152$) might indicate that perhaps a few antigenic sites are β (i.e., extended)-structures. This hypothesis is bolstered by considering the strong anti-correlation of α- and β-propensities ($\bar{r} = -0.369$ and $r_0 = -0.646$). Although most antigenic sites take α-conformations (7), β-Propensity still retains some significance, so some antigenic sites probably do assume β-conformations. There is also an unusual correlation of β-propensity with turn and coil propensities in the antigenic sites ($p = 0.041$ and 0.013). The correlation significances are difficult to explain without the presence of β-conformations. Single β-strands are unstable, however, so the significances are explicable if the extended conformations stabilize as at least two β-strands joined by a coil or turn. The absence of β-amphipathicity ($p = 0.855$) is similarly explained because the coil or turn, being of uncertain length, destroys the two-residue periodicity required by a significant β-amphipathicity. Also, β-structures are often twisted, again reducing the required periodicity. β-amphipathicity would help to stabilize a β-sheet at an aqueous interface and may well be present, although undetectable by our techniques.

The anti-significance of coil propensity is crucial to any theory that immunodominant sites tend to have stable conformations ($p = 0.976$, i.e., $1\text{-}p = 0.024$). If peptide conformation tends to be more random when the peptide has coil propensity, the theory would be much weaker without this anti-significance. The significant correlation of coil and β-propensities suggests that the few coils occurring in the antigenic peptides join β-strands.

Our results indicate that T cells tend to recognize the parts of a foreign molecule which have the greatest propensities to α-, and perhaps β-, conformations. The propensities of T cell antigenic sites are perhaps quite difficult for an organism to change because they may correlate with structure-function properties of proteins. In this respect, teliologically, the T cell recognition system would appear an excellent complement to the antibody system which recognizes the surface features of a protein. If the structure-function supposition is correct, the T cells must of necessity recognize *both* α- and β-conformations, because otherwise evolution would favor the unrecognized structure to evade detection.

Segmental amphipathicity. Table II shows that segmental amphipathicity is not correlated with the experimentally determined antigenic sites ($p = 0.843$ and 0.887). Our statistics consistently confirmed the absence of segmental amphipathicity in antigenic sites. Because segmental amphipathicity has enjoyed unwarranted currency n the basis of anecdotal evidence, we discuss it at length. Some possible arguments to justify the lack of statistical significance follow.

The statistics chosen may not reflect the property "segmental amphipathicity." This argument can be raised against any statistic purporting to represent a property. differential hydrophobicity is the only previous attempt to give this property a quantitative operational definition (7). Any operati nal definition can be checked for statistical significance: maximal differential hydrophobicity, systematizing the procedure of Corradin et al. (10), is another such definition. Neither definition showed statistical significance.

If segmental amphipathicity is present only in a subset of the antigenic sites, the remaining sites might mask its statistical significance. Those sites in hen lysozyme for which the segmental amphipathicity hypothesis was first invoked (7, 13) might then be this subset, or perhaps the subset examined by Corradin et al. (10). Several different subsets of the antigenic sites, including the two mentioned, have been tested on their own for segmental amphipathicity; we also tried several different hydrophobicity scales, including the Kyte-Doolittle (47) scale that Corradin et al. (10) used: segmental amphipathicity was, again, not significant.

We have been most thorough in trying to detect segmental amphipathicity. Although no statistic can ever rule out individual exceptional circumstances, *our statistical criteria give absolutely no support to the segmental amphipathicity hypothesis.* Where applicable in protein and DNA research, the statistical method of this study provides an objective criterion for determining significance. The case of segmental amphipathicity shows that anecdotal evidence can be very misleading.

The absence of segmental amphipathicity is also consistent with the conformational hypothesis. Assume, for the sake of argument, that a T cell antigen is recognized while it is in the amphipathic environment overlying a shallow hydrophobic protein cleft. Because a segmentally amphipathic peptide cannot penetrate far into the hypothetical cleft, any attempt to form regular internal hydrogen bonds leads to exposure of hydrophobic residues. This effect produces conformational destabilization. Segmental amphipathicity therefore runs counter to the conformation hypothesis *if* antigenic stabilization takes place in such a hypothetical shallow cleft.

These studies have brought to light a number of properties associated with immunodominant antigenic sites for helper T cells. These may be useful in the rational design of synthetic vaccines. It should be cautioned, however, that in any given individual, only a subset of such sites will be seen, depending on MHC-linked immune response genes, self-tolerance to homologous self-antigens, and other genetic and environmental constraints of the host (3–5, 14).

Acknowledgment. We thank Dr. David Alling for critical reading of the manuscript and helpful discussion.

REFERENCES

1. **Streicher, H. Z., I. J. Berkower, M. Busch, F. R. N. Gurd, and J. A. Berzofsky.** 1984. Antigen conformation determines processing requirements for T-cell activation. *Proc. Natl. Acad. Sci. USA 81:6831.*
2. **Unanue, E. R.** 1984. Antigen-presenting function of the macrophage. *Annu. Rev. Immunol. 2:395.*
3. **Benacerraf, B.** 1978. A hypothesis to relate the specificity of T lymphocytes and the activity of I region-specific *Ir* genes in macrophages and B lymphocytes. *J. Immunol. 120:1809.*
4. **Rosenthal, A. S.** 1978. Determinant selection and macrophage function in genetic control of the immune response. *Immunol. Rev. 40:136.*
5. **Berzofsky, J. A.** 1980. Immune response genes in the regulation of mammalian immunity. In *Biological Regulation and Development. Vol. 2.* R. F. Goldberger, ed. Plenum Press, New York. Pp. 467–594.
6. **Segrest, J. P., and R. J. Feldmann.** 1977. Amphipathic helixes and plasma lipoproteins: a computer study. *Biopolymers 16:2053.*
7. **DeLisi, C., and J. A. Berzofsky.** 1985. T-cell antigenic sites tend to

be amphipathic structures. *Proc. Natl. Acad. Sci. USA 82:7048.*

8. **DeLisi, C., J. Cornette, H. Margalit, K. Cease, J. Spouge, and J. A. Berzofsky.** 1986. The role of amphipathicity as an indicator of T cell antigenic sites on proteins. In *The Immunogenicity of Protein Antigens: Repertoire and Regulation.* E. E. Sercarz and J. A. Berzofsky, ed. CRC Press, Boca Raton. In press.
9. **Berzofsky, J. A., J. Cornette, H. Margalit, I. Berkower, K. Cease and C. DeLisi.** 1986. Molecular features of Class II MHC-restricted T cell recognition of protein and peptide antigens: the importance of amphipathic structures. *Curr. Top. Immunonol. 130:14.*
10. **Corradin, G. P., C. J. A. Wallace, A. E. I. Proudfoot, and S. Baumhuter.** 1986. Murine T cell response specific for cytochrome c. In *The Immunogenicity of Protein Antigens: Repertoire and Regulation.* E. E. Sercarz and J. A. Berzofsky, ed. CRC Press, Boca Raton. In press.
11. **Pincus, M., F. Gerewitz, R. H. Schwartz, and H. A. Scheraga.** 1983. Correlation between the conformation of cytochrome c peptides and their stimulatory activity in a T-lymphocyte proliferation assay. *Proc. Natl. Acad. Sci. USA 80:3297.*
12. **Schwartz, R. H., B. S. Fox, E. Fraga, C. Chen, and B. Singh.** 1985. The T lymphocyte response to cytochrome c. V. Determination of the minimal peptide size required for stimulation of T cell clones and assessment of the contribution of each residue beyond this size to antigenic potency. *J. Immunol. 135:2598.*
13. **Allen, P. M., D. J. Strydom, and E. R. Unanue.** 1984. Processing of lysozyme by macrophages; identification of the determinant recognized by two T cell hybridomas. *Proc. Natl. Acad. Sci. USA. 81:2489.*
14. **Berkower, I. J., H. Kawamura, L. A. Matis, and J. A. Berzofsky.** 1985. T cell clones to two major T cell epitopes of myoglobin: effect of I-A/I-E restriction on epitope dominance. *J. Immunol. 135:2628.*
15. **Berzofsky, J. A.** 1985. The nature and role of antigen processing in T cell activation. In *The Year in Immunology 1984–1985.* J. M. Cruse and R. E. Lewis, ed. Karger, Basel. Pp. 18–24.
16. **Chou, P. Y., and G. Fasman.** 1974. Prediction of protein conformation. *Biochemistry 13:222.*
17. **Jernigan, R. L., and S. C. Szu.** 1979. Conformational energy minimization in the approximation of limited range interactions. *Macromolecules 12:1156.*
18. **Wada, A., and H. Nakamura.** 1981. Nature of the charge distribution in proteins. *Nature 293:757.*
19. **Hol, W. G. J., L. M. Halie, and C. Sander.** 1981. Dipoles of the α-helix and β-sheet: their role in protein folding. *Nature 294:532.*
20. **Livingstone, A., and C. G. Fathman.** 1987. The structure of T cell epitopes. *Annu. Rev. Immunol.* In press.
21. **Cease, K. B., I. Berkower, J. York-Jolley, and J. A. Berzofsky.** 1986. T cell clones specific for an amphipathic alpha helical region of sperm whale myoglobin show differing fine specificities for synthetic peptides: a multi-view/single structure interpretation of dominance. *J. Exp. Med. 164:1780.*
22. **Berkower, I. J., G. K. Buckenmeyer, and J. A. Berzofsky.** 1986. Molecular mapping of a histocompatibility-restricted immunodominant epitope with synthetic and natural peptides: implications for T cell antigenic structure. *J. Immunol. 136:2498.*
23. **Hansburg, D., T. Fairwell, R. H. Schwartz, and E. Appella.** 1983. The T cell response to cytochrome *c*. IV. Distinguishable sites on a peptide antigen which affect antigenic strength and memory. *J. Immunol. 131:319.*
24. **Hackett, C. J., B. Dietzschold, W. Gerhard, B. Ghrist, R. Knorr, D. Gillesaen, and F. Melchers.** 1983. Influenza virus site recognized by a murine helper T cell specific for H1 strains. *J. Exp. Med. 158:294.*
25. **Hurwitz, J. L., E. Heber-Katz, C. J. Hackett, and W. J. Gerhard.** 1984. Characterization for the murine T_H response to influenza virus hemagglutinin: evidence for three major specificities. *Immunology 133:3371.*
26. **Lamb, J. R., D. D. Eckels, P. Lake, J. N. Woody, and N. Green.** 1982. Human T-cell clones recognize chemically synthesized peptides of influenza hemagglutinin. *Nature 300:66.*
27. **Thomas, J. W., W. Danho, E. Bullesbach, J. Fohles, and A. S. Rosenthal.** 1981. Immune response gene control of determinant selection. III. Polypeptide fragments of insulin are differentially recognized by T but not by B cells in insulin-immune guinea pigs. *J. Immunol. 126:1095.*
28. **Allen, P. M., D. J. McKean, B. N. Beck, J. Sheffield, and L. H. Glimcher.** 1985. Direct evidence that a Class II Molecule and a simple globular protein generate multiple determinants. *J. Exp. Med. 162:1264.*
29. **Shastri, N., A. Oki, A. Miller, and E. E. Sercarz.** 1985. Distinct recognition phenotypes exist for T cell clones specific for small peptide regions of proteins. Implications for the mechanisms underlying major histocompatibility complex-restricted antigen recognition and clonal deletion models of immune response gene defects. *J. Exp. Med. 162:332.*
30. **Katz, M. E., R. M. Maizels, L. Wicker, A. Miller, and E. E. Sercarz.** 1982. Immunological focusing by the mouse major histocompatibility complex: mouse strains confronted with distantly related lysozymes confine their attention to very few epitopes. *Eur. J. Immunol. 12:535.*
31. **Shimonkevitz, R., S. Colon, J. W. Kappler, P. Marrack, and H. Grey.** 1984. Antigen recognition by H-2-restricted T cells. II: A tryptic ovalbumin peptide that substitutes for processed antigen. *J. Immunol. 133:2067.*
32. **Milich, D. R., G. B. Thornton, A. McLachlan, M. K. McNamara, and F. V. Chisari.** 1986. T and B cell recognition of native and synthetic pre-S region determinants on HBsAg. In *Modern Approaches to Vaccines.* R. Chanock, R. A. Lerner, and F. Brown, eds. Cold Spring Harbor Laboratories, New York. In press.
33. **Francis, M. J., C. M. Fry, D. J. Rowlands, F. Brown, J. L. Bittle, R. A. Houghten, and R. A. Lerner.** 1985. Immunological priming with synthetic peptides of foot and mouth disease virus. *J. Gen. Virol. 66:2347.*
34. **Corradin, G. P., M. A. Juillerat, C. Vita, and H. D. Engers.** 1983. Fine specificity of a BALB/c T cell clone directed against beef apo cytochrome *c*. *Mol. Immunol. 20:763.*
35. **Milich, D. R., D. L. Peterson, G. G. Leroux-Roels, R. A. Lerner, and F. V. Chisari.** 1985. Genetic regulation of the immune response to Hepatitis B surface antigen (HBsAg). VI. Fine specificity. *J. Immunol. 134:4203.*
36. **Guillet, J.-G., M.-Z. Lai, T. J. Briner, J. A. Smith, and M. L. Gefter.** 1986. The interaction of peptide antigens and Class II major histocompatibility antigens as studied by T-cell activation. *Nature.* In press.
37. **Macfarlane, R. L., B. Dietzschold, T. J. Wiktor, M. Kiel, R. Houghten, R. A. Lerner, J. G. Sutcliffe, and H. Koprowski.** 1984. T cell responses to cleaved rabies virus glycoprotein and to synthetic peptides. *J. Immunol. 133:2748.*
38. **Hammersley, J. M., and D. C. Handscombe.** 1964. *Monte Carlo Methods.* Methuen, London.
39. **Dwass, M.** 1970. *Probability and Statistics.* Benjamin, New York. P. 510.
40. **Fauchère, J. L., and V. Pliska.** 1983. Hydrophobic parameters II of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem. 18:369.*
41. **Garnier, J., D. J. Osguthorpe, and B. Robson.** 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol. 120:97.*
42. **Mayer, J. E., and M. G. Mayer.** 1940. *Statistical Mechanics.* Wiley and Sons, New York. P. 56.
43. **Kabsch, W., and C. Sander.** 1983. How good are predictions of protein secondary structure? *FEBS Lett. 155:179.*
44. **Naquet, P., M. L. Phillips, J. Ellis, R. Hodges, B. Singh, and T. L. Delovitch.** 1986. *Immunogenicity of Proteins: Repertoire and Regulation.* E. E. Sercarz and J. A. Berzofsky, eds. CRC Press, Boca Raton.
45. **Robson, B.** 1974. Analysis of the code relating sequence to conformation in globular proteins. *Biochem. J. 141:853.*
46. **Barrett, A. J., ed.** 1977. *Proteinases in Mammalian Cells and Tissues, Research Monographs in Cell and Tissue Physiology. Vol. 2.* J. T. Dingle, general editor. North Holland, New York.
47. **Kyte, J., and R. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157:105.*